MEDIA COVERAGE OF EDUCATIONAL TESTING: UNDERSTANDING ISSUE DIMENSIONS USING TOPIC MODELING

John Wachen

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Education (Policy, Leadership, and School Improvement)

Chapel Hill
2017

Approved by:

Lora Cohen-Vogel

Frank Baumgartner

Gregory Cizek

Eric Houck

Kirsten Kainz

ABSTRACT

John Wachen: Media Coverage of Educational Testing:
Understanding Issue Dimensions Using Topic Modeling
(Under the direction of Lora Cohen-Vogel)

This study examined media coverage of testing in schools in two newspapers – the *New York Times* and *Education Week* – over the 20-year period from 1996 to 2015. Using a conceptual framework informed by framing theory and a quantitative text analysis technique called structural topic modeling on a dataset of over 8,000 articles, this study identified shifts in topic coverage and traced the evolution of the framing of the issue in the media across time. Findings indicate a shift in framing from a more positive portrayal of testing in the years prior to passage of the federal No Child Left Behind Act to a more negative portrayal in the years after passage of the law. The study contributes to the literature on media framing of education issues and the emerging literature on applications of topic modeling techniques to education policy.

ACKNOWLEDGEMENTS

Graduate school is both a privilege and a demanding intellectual journey. It would be unwise and foolhardy to embark on this journey without help along the way from friends, family, and colleagues. My development as a scholar is thanks to the following people and many more.

To my advisor, Lora. Thank you for guiding me through the steps of graduate school. You are a steadfast champion for all of your students and you motivated me to strive to be the best student and scholar I could possibly be in every respect.

To my committee, Lora, Frank Baumgartner, Greg Cizek, Eric Houck, and Kirsten Kainz. You each played an integral part in my development as a scholar. Thank you for your insights, your suggestions, and your support of my research. I will always be proud to see your names on the title page of this work alongside mine.

To my wonderful family, thank you thank you thank you. Mom and Dad, your quiet, unwavering support and belief in me during all of my professional and scholarly endeavors is the greatest gift you could possibly give. Matt, Jen, Will, and Ollie, when I am around you all, I truly feel like a kid again.

To my two best comrades at UNC, Cheryl and Mark. Mark, you always listened to me vent about the ups and downs of graduate school and Carolina sports. Go Heels. Cheryl, I doubt very much that I would have finished my doctorate without your support through those years. My entire experience in Chapel Hill was infinitely better with you as my friend. Thank you.

Finally, to Kim. Thank you for your smile, your laughter, and your love while I was finishing this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The use of assessments in public schools to gauge student achievement and make comparisons across schools, districts, and states has a long history in the United States. Horace Mann and his colleagues started the testing movement in Massachusetts public schools in the 1840s, comparing Boston schools to those in nearby suburbs. The reaction was similar to what has transpired in the testing debate ever since: scores were lower than expected, the tests were attacked as being invalid, and policymakers debated holding schools and teachers accountable for poor performance (Reese, 2013).

Testing has long been a feature of public schooling in America. In recent decades, high stakes have increasingly been attached to these tests. For example, high school graduation exit exams, grade retention policies in elementary schools based on reading proficiency, and teacher performance and pay metrics based on test scores are reform efforts based on high-stakes testing that have been implemented in recent decades (McIntosh, 2012; Springer et al., 2012; Workman, 2014).

Why Study Testing in Schools?

The passage of the No Child Left Behind Act (NCLB) of 2001 signaled an important change in the federal government's role in education. For the first time, schools and districts nationwide faced the possibility of sanctions based on student performance on tests. NCLB initially mandated annual testing in reading and math in grades 3 through 8 and once in high school. Beginning in the 2007-2008 academic year, states were also required to test in science at least once in three different grade spans: grades 3-5, grades 6-9, and grades 10-12. Many states

had accountability policies and testing systems in place prior to NCLB, effectively setting the stage for a new federal role in elementary and secondary education (Manna, 2006; McDonnell, 2005; National Conference of State Legislatures, 2005). The law solidified test-based accountability as a core feature of education reform and expanded the federal government's role. In the NCLB era, high-stakes testing became an ingrained part of the culture of public schools. In the years since passage of the law, however, there has been an increasingly organized opposition to testing of students. Recent examples of state and federal legislative action and the growing "opt out" movement among parents and educators illustrate this growing backlash against testing (Chandler, 2014; Fiedler, 2014; Hernandez, 2014; National Education Association, 2013).

An enduring narrative of American public schools since the publication of the federal report *A Nation at Risk* in 1983 is that public schools have largely failed to provide a high quality education to many students and that American students are falling behind their counterparts in many other economically advanced countries around the world (Chubb, 1988; Evers & Walberg, 2004; Klein, 2011; Lynch, 2015; Vockell, 1993). Several metrics have been used to indicate the extent to which the quality of American public schooling needs to be improved. For example, the number of schools failing to make adequate yearly progress under NCLB was often cited as evidence. In a 2008 report, for example, researchers found that 24,200 schools (approximately 26% of all public schools) failed to make adequate yearly progress in 2004-2005 (Stullich, Eisner, & McCrary, 2008). The high number of students in need of remediation upon matriculation in postsecondary institutions is often cited as another indication of the need to reform K-12 schooling. Low graduation rates in some schools and districts is another metric, as is student performance on national and international tests such as the National Assessment of

Educational Progress (NAEP) or the Program for International Student Assessment (PISA). In recent years, for example, student performance on the NAEP has stagnated or declined slightly, a development that has not gone unnoticed in the media (Brown, 2015; Rich, 2015; Summers, 2014).

These different indicators are often cited by policymakers and scholars in reference to the substandard state of public schooling in the country. An emblematic sentiment, based on standardized test score results, is provided by Evers and Walberg (2004):

> Though the achievement of American students is comparable with that of students in other countries when American students begin school, they fall increasingly behind as they progress through the grades. By the end of high school, their achievement is near the bottom of advanced countries, despite American schools' being close to the top in per-student spending among economically advanced countries (p. viii-ix).

These various indicators have led to persistent calls to reform public schooling across the country. The approach that has gained popularity in recent decades, but has a long history as a part of schooling, is the use of test-based accountability as a reform strategy to influence students and educators toward better performance. After years of policymakers at the state and federal level developing policies that include high-stakes testing (tests tied to accountability and with important consequences) as a core component of reform strategies, a culture of testing is now well established in public education (McDonnell, 2008).

The Culture of Testing

Scholars have noted that test-based accountability is now the dominant education policy in the nation (McDonnell, 2008; Ravitch, 2010). For example, an analysis of presidential discourse on testing found that George H. W. Bush, Bill Clinton, and George W. Bush all

advocated for test-based accountability as a policy solution for various problems with public schooling in the U.S. (Wachen, 2014). But how prominent is testing in schools? There has been relatively little data collected on how much testing is occurring in public schools across the nation (Hart et al., 2015). NCLB mandated that states conduct annual testing of students in math and reading in grades 3 through 8 and once in high school, and this federal requirement remains in place with the reauthorization of NCLB as the Every Student Succeeds Act (ESSA). A high-profile report by the Council of the Great City Schools indicated that students in 66 urban districts across the country were required to take an average of 112 tests from pre-K to grade 12, although many of these tests were not coupled with accountability (Hart et al., 2015). According to another recent study, the average amount of time students spend on testing in a sample of a dozen urban districts was 1.7 percent of the school year (Teoh, Coggins, Guan, & Hiler, 2014). However, as critics have noted, this percentage does not account for hours of test preparation, reporting, and other activities related to testing, which is more difficult to quantify (Keeling, 2014). Regardless of the amount of time that testing takes up in a school year or a student's academic career, test-based accountability has become a core policy idea in American public schooling. However, recent developments suggest that there may be a shift occurring in the dominance of this policy idea.

The Recent Shift

Several developments in recent years suggest that there is a shift in thinking and an increasingly strong challenge to the narrative among policymakers and the public that high-stakes testing is a positive and necessary aspect of school reform efforts. These developments have occurred at local, state, and national levels. At the local level, the first development is the growing opt-out movement, a movement that first gained momentum in 2014 and has seen

increasing numbers of families keeping their children out of school on days when tests are administered as a form of passive protest (Schweig, 2016). Some estimates suggest that about 670,000 students opted out of standardized tests in 2015 (FairTest, 2016). Although this is a very small fraction of the approximately 50 million students in public elementary and secondary schools, the movement has gained enough momentum to be addressed by the federal government in ESSA. ESSA includes language that states can establish opt out policies for parents for federally required tests. However, the law also retains the requirement (initially established as part of NCLB) that 95 percent of all students must participate in the tests (Every Student Succeeds Act, 2015).

The second development is state legislative action pushing back against the proliferation of high-stakes testing. Some recent examples of state legislative efforts illustrate this growing backlash. In New York City, for example, after 12 years of focusing on testing and accountability in the city's public schools under Mayor Michael Bloomberg, the New York City Department of Education announced in April 2014 that it was reducing the role of standardized tests in determining admission to selective schools and determining which students to hold back (Hernandez, 2014). Similarly, after two decades of developing a statewide accountability system with end-of-year standardized tests at its core, Virginia legislators passed a bill in April 2014 reducing the number of tests that elementary and middle school students must take (Chandler, 2014).

A similar development is occurring at the federal level, where there is a shift in the federal government's position on testing amid more vocal opposition to high-stakes testing from national teachers unions and other interest groups. After decades of support for test-based accountability in schools, the federal government has, over the past several years, begun to

backtrack on this issue. In 2014, for example, Secretary of Education Arne Duncan stated that "the sheer quantity of testing - and test prep - has become an issue" and went on to announce that the U.S. Department of Education wanted to be part of the solution to this problem (Gewertz, 2014). In March 2014, U.S. Representatives introduced a bill titled "Student Testing Improvement and Accountability Act" that would reduce the number of federally mandated tests from fourteen to six. In 2015, President Obama and the U.S. DOE announced a "Testing Action Plan," an initiative to reduce unnecessary testing and create fewer, higher quality tests (U.S. Department of Education, 2015a).

In spite of this apparent shift, there are several factors that might suggest that support for high-stakes testing would remain high. They include the increasingly active involvement in education by the federal government, which has traditionally supported test-based accountability in schools (although the federal role in education may be tempered somewhat through the implementation of the Every Student Succeeds Act, which was passed in late 2015) (Orfield, 2016); the continuing narrative of failing public schools and American students' inability to keep up with international peers (Jennings, 2016); and the dominant focus of many school reform efforts on data-driven decision making, which is strongly linked to standardized testing. However, developments such as the opt-out movement and federal and state legislative efforts to limit testing suggest that the American public appears to be increasingly exposed to messages about potentially negative consequences of an emphasis on testing in schools. But to what extent does media coverage play a role in shaping public perceptions and policy decisions about testing in schools?

Research shows that media play an important role in shaping public policy and public perceptions yet there are very few studies examining the role of media in shaping education

policy specifically and none analyzing the role of media in influencing the testing debate. Given the role that media play in shaping policy and public agendas and the recognition that media are not objective conduits of information (Shanahan et al., 2008), it is necessary for education researchers and policymakers to understand how media outlets frame and recast education issues. Despite the dominance of high-stakes testing in education policy in the U.S., there has not yet been a systematic or empirical review of media coverage of testing issues. Additionally, we know little about what preceded or prompted the recent shift in perceptions about testing or whether and how media coverage of the issue contributed to this shift.

What are High-Stakes Tests?

What do we mean when we use the term high-stakes test? Technically, tests themselves are not high stakes. High-stakes testing actually means the use of tests for high-stakes decisions or for accountability purposes (Kober, 2015). That is, when important or consequential decisions are attached to tests, they become high stakes. For example, a driver's license test would be considered high stakes because it determines whether or not the test taker is legally allowed to drive a vehicle. In terms of education, the Glossary of Education Reform defines a high-stakes test as, "Any test used to make important decisions about students, educators, schools, or districts, most commonly for the purpose of accountability" (Great Schools Partnership, n.d.). These important decisions often come in the form of rewards or sanctions. Other definitions emphasize that high-stakes testing refers not only to rewards or sanctions, but also a connection to an underlying intended outcome: ensuring change or reform (McDonnell, 2005).

This conceptualization of testing as high stakes and driving decisions may differ from the stated purposes and designs of some tests. It can be problematic when interpretations of test scores or actions based on test scores are not evidence-based. That is, when actions based on the

interpretation of test scores are not supported, validity becomes an issue. Validity is the degree to which interpretations and inferences of test scores are supported by evidence. Interpretations that "make sense and are supported by appropriate evidence are considered to have high validity" (Kane, 2013, p. 3). Another potential issue occurs when tests are used for multiple purposes, such as both accountability purposes and student-level diagnostic purposes (Miller, 2008). For example, educational benchmarking tests are designed to assess academic skills and knowledge acquired through schooling or diagnose areas of student difficulty. Diagnostic results from these tests can then inform curriculum and instruction with the goal of improving teaching and learning (Miller, 2008). However, if scores from benchmarking tests such as these were also used to evaluate teachers or schools, a disparity might exist if evidence does not support the application of test score data for this other purpose, leading to validity problems.

According to Stobart and Eggen (2012), there are three broad categories of uses of tests when they become high stakes: selection, placement, and certification; raising standards; and accountability. Examples of high-stakes tests for selection, placement, and certification include the test that is arguably the nation's most famous, the SAT; the ACCUPLACER, which determines the college readiness of incoming students; and Praxis tests for teacher certification and licensure (Jennings, 1998; Lemann, 1999). Examples of tests for raising standards include the recently developed Partnership for Assessment of Readiness for College and Careers and Smarter Balanced tests that are aligned with the Common Core State Standards. High-stakes tests for accountability include end-of-grade exams for promotion to the next grade level, high school exit exams, and annual tests to determine yearly progress in accordance with federal policy.

When tests are used for accountability purposes, the appropriate use of tests becomes critically important, as do issues of test validity and reliability. The alignment of the purpose and the use of tests takes on additional significance when students, teachers, or schools are held accountable and potentially rewarded or penalized based on test results. The ACCUPLACER test, for example, is designed to help "high school and college educators advise students about course selection, preparation, and opportunities for success" (College Board, n.d.). Therefore, when the test is used to make decisions about incoming college student course placements and remediation needs, purpose is aligned with use. In contrast, if the ACCUPLACER were used to make decisions about the effectiveness of high school principals, purpose and use would be misaligned.[1] Concerns about the use of tests for accountability purposes were prevalent even before the passage of NCLB, as a National Research Council report from 1999 illustrates:

> The use of tests in school reform raises difficult questions in relation to so-called high-stakes consequences for students—that is, when an individual student's score determines not just who needs help but whether a student is allowed to take a certain program or class, or will be promoted to the next grade, or will graduate from high school (Heubert & Hauser, 1999, p. 14).

When used in this way, tests can become a highly controversial and politicized aspect of public education. I discuss the notion of tests as policy instruments as well as the controversies associated with this use in greater detail in Chapter Two.

[1] Several studies suggest that there is limited predictive validity of placement exam scores from tests such as ACCUPLACER and that using multiple measures for placement decisions could reduce misplacement (Belfield & Crosta, 2012; Scott-Clayton, 2012).

High Stakes for Whom?

In addition to the three types of uses of tests outlined by Stobart and Eggen (2012), decisions based on these tests can impact different stakeholders in education, including students, educators, schools, and politicians. Perhaps the most obvious of these is the impact on students. Many of these tests directly impact student achievement and progression through the education system through retention or graduation policies or selection processes into college, all of which are often tied to test scores. Tests can also be high stakes for educators. For example, value-added models use data from standardized tests to determine teachers' and administrators' contributions to student learning and these models are typically implemented in an attempt to identify educators who are particularly effective or ineffective (AERA, 2015). Value-added models have been incorporated by districts and states into evaluation systems used for making personnel decisions. Tests can also be high stakes for schools. Tests used to determine adequate yearly progress of schools can impact federal funding and also result in corrective actions, including replacing school leadership or restructuring the school itself. A potentially perverse consequence of the high-stakes attached to testing is that some schools may feel pressure to doctor test results given the possibility of these corrective actions.

Less directly, stakes can also be high for politicians. Results from international tests such as PISA or TIMSS receive attention in the media, often comparing U.S. student performance with the performance of students in other participating countries. These tests may at first appear to be low stakes, since individual students and schools are not impacted by the results. However, with the growing importance of global economic competitiveness and the central role of education in international comparisons, these tests have become a high-stakes phenomenon for politicians and governments (Stobart & Eggen, 2012). In an examination of the response of

several countries to PISA results in 2000, for example, Baird et al. (2011) found that Norway experienced a "PISA shock" when results were lower than neighboring countries, particularly given generous spending on education. The surprising results had political implications and led to large scale curricular and accountability reforms (Baird et al., 2011; Sjøberg, 2007).

As this discussion makes clear, tests can be used for many different purposes and many tests have stakes attached. It is therefore difficult to disaggregate coverage of tests with high stakes attached from coverage of testing broadly, especially given that the stakes are not necessarily direct, as is the case for politicians. More importantly, for the purposes of the current study, in which I am interested in capturing all aspects of the debate over testing in media coverage to better understand how testing is discussed and portrayed, it would not be productive or beneficial to attempt to capture only articles that explicitly discuss testing that is high stakes. Instead, I am interested in explicating the framing of the issue, which can best be done by an examination of coverage of testing that is as comprehensive as possible. Therefore, I conducted a broad search of newspaper articles on testing (described in detail in Chapter Three) in order to cast a wide net and capture as many articles about the testing debate in media coverage as possible. This was necessary to gain a greater understanding of as many aspects of the debate as possible.

One final note. The terminology of testing can be confusing. In an article on unaddressed issues related to classroom assessment, Cizek (2000) distinguishes between assessment (which is similar to, but can also be broader than, the term testing) and evaluation. Assessment is the "planned observation and collection of information" for purposes such as identifying students' strengths or areas of weakness. Evaluation is the process of ascribing value to the results of the collection of information (p. 16). For the purposes of this study, I am

interested in assessment, not evaluation, but I use the terms "test" and "testing" rather than "assessment" as those are the terms commonly used in media coverage of the issue.

In this dissertation, I explore shifts in the debate over the past 20 years by empirically identifying the arguments for and against high-stakes testing in media coverage of the issue and tracing the evolution of its framing. Because the purpose of this study is to unpack the ways in which high-stakes testing has been framed over time, the study is guided by the following research questions:

1. How has the issue of testing in schools been framed in media coverage?
2. How has the debate over testing evolved over time?
3. To what extent do certain dimensions of the issue dominate the debate at different periods of time?
4. Do the frames in media coverage (and shifts in framing over time) differ in general versus professional newspapers?

In Chapter Two, I outline the history of testing in schools; discuss the popularity of testing and test-based accountability as a policy tool; and describe the conceptual framework used in the study, which draws on scholarly work from political science and communication on framing. In Chapter Three, I describe the methodological approach used to study media coverage of testing and review the application of text analytics to political science and education research.

CHAPTER 2: LITERATURE REVIEW

History of Testing

Early History

The testing of students has a long history in public education in the United States. Scholars have analyzed various developments in the history of public schooling that have contributed to the growth of the testing movement, including the mid-nineteenth century debate about the quality of public schools in Boston (Reese, 2013), the influence of intelligence testing and psychology in the early twentieth century (Giordano, 2005), and the entrenchment of the high-stakes testing movement in the early twenty-first century (Ravitch, 2010).

From the beginning, testing was controversial – its merits and deficiencies often being debated in public discourse. The Boston public school system was the site of the first real testing controversy in 1845 (Reese, 2013). Media coverage played a large part in the controversy, both by reporting on test score results (often highlighting what appeared to be shockingly dismal student performance) and by publishing editorials and letters to the editor supporting or condemning testing (Reese, 2013). Many of the issues related to testing raised during the nineteenth century bear an uncanny resemblance to the arguments in the testing debate today.

In addition to being a historically controversial issue, testing has always been a highly political issue as well. The "testing wars" that played out in Boston in the mid-19th century demonstrate that tests were seen as more than simply providing information on students' levels of content mastery. Test scores were interpreted as measures of school and district quality, as well as teacher and administrator quality. Test scores were also connected to community pride.

The Boston case illustrates that some stakeholders were adamantly against publishing test scores because they might tarnish the image of what was considered at the time to be the nation's best school system (Reese, 2013). Additionally, test scores were seen as an indicator of educational productivity (i.e. the achievement produced by a school or district relative to its spending). Newspapers in Boston continually reported on the testing controversy that unfolded in that city throughout 1845 (Reese, 2013).

Media coverage continued to highlight test results in later decades, as illustrated by an article in the November 1880 issue of *Harper's New Monthly Magazine* that described the "startling" results of examinations in a suburban Massachusetts school district which revealed that "the scholars of fourteen years of age did not know how to read, to write, or to cipher" and discussed the implications of the poor performance by students on the examination (Adams, 1880).

In the early 1900s, the 'new science of testing' began to develop through the work of Edward Thorndike and others. This new science was principally based on psychological testing and, as such, testing remained mostly an area of special expertise of researchers and psychologists, with little interest or knowledge about testing among the public (Brookhart, 2013). Scholars have written extensively about the development of the field of psychological testing through U.S. Army intelligence testing, particularly during World War I, when it became necessary to sort millions of recruits into various roles and positions (Giordano, 2005; Gould, 1996; Gregory, 2013; Lemann, 1999). The development of the National Assessment of Educational Progress (NAEP) in the late 1960s was a significant development contributing to the movement toward test-based accountability. NAEP is intended to provide a nationally representative measure of what U.S. students know in a variety of subject areas. Similar to many

state administered standardized tests, NAEP reports the percentage of students meeting proficiency in subjects including math, science, reading, writing, civics, geography, and history. Results from NAEP often reveal substantially lower levels of proficiency than state test results show. This inconsistency has led to questions about the rigor of state standards and tests and has been used by federal policymakers to advocate for national standards (Jennings, 1998).

The Growth of High-Stakes Testing

Although the history of testing in American schools stretches back over 150 years, the movement to link tests to accountability policies grew substantially in more recent decades. Linn (2000) identified 5 waves of growth of accountability testing, roughly mapping onto the five decades from the 1950s to 2000. Scholars have noted that as this movement progressed, testing grew into the central element of accountability systems. Dwyer (2004), for example, claimed that the movement toward education reform policies that relied heavily on high-stakes tests really started to take off in the 1990s. Similarly, Carnoy and Loeb noted in 2002 that test scores were "rapidly becoming the end-all of state accountability reforms" (p. 307). In the following sections, I briefly cover the recent history of the growth in high-stakes testing. This history is not intended to be comprehensive. A full accounting of how test-based accountability became a dominant policy idea in education is beyond the scope of this dissertation. Numerous books have been written on various aspects of the history of testing in the United States (Lemann, 1999; Ravitch, 2010; Reese, 2013). My goal is to broadly review the recent history of testing and to specifically provide context for the time period covered in my analysis, which spans the 20-year period from 1995 to 2015.

Program accountability and the minimum competency movement. In the early and middle years of the 20th century, achievement tests were primarily used by schools and teachers

to gauge student learning and provide feedback on instruction (Behuniak, 2003). With growing concern over declining test scores and the perception that many students were being moved along from grade to grade without sufficient academic skills, state-level policymakers and education reformers turned to minimum competency testing starting in the 1960s (Haertel & Herman, 2005). Based on an established basic skills standard, minimum competency tests were used to determine grade promotion. In some states, passing the test was a requirement to obtain a high school diploma. Scholars have suggested that *minimum* competency resulted in required proficiency levels that were very low, rendering the standards ineffectual at catalyzing changes to curriculum or instruction (Haertel & Herman, 2005). Effectiveness aside, these tests set the stage for further test-based accountability reform efforts.

Standards-based movement. The 1980s and 1990s signaled a shift in policies and approaches to education reform and a new focus on high-stakes testing that can be attributed to a variety of factors. Scholars have consistently noted that the publication of *A Nation at Risk* by Ronald Reagan's National Commission on Excellence in Education in 1983 was a major turning point in education, both in terms of how public schooling was perceived and in terms of the need for, and approaches to, education reform (Giordano, 2005; Kean, 2003; Mehta, 2015; Ravitch, 2010). *A Nation at Risk* was a politically charged, incendiary document about the declining quality of education in America, especially when compared to other countries. In particular, poor performance on tests and structural changes to the U.S. economy resulted in pressure from the business community and others to increase productivity in public education. The authors of *A Nation at Risk* recommended implementing educational standards and standardized testing at the state and local levels and called for an end to the minimum competency testing movement (National Commission for Excellence in Education, 1983). In response to the inflammatory

rhetoric of *A Nation at Risk*, the National Governors Association issued *Time for Results* (National Governors Association, 1986), which called for greater accountability in public schools and standards-based reform. Education historian Diane Ravitch argued that *A Nation at Risk* catalyzed the standards movement of the 1980s and 1990s but when that movement failed to gain substantial traction, the reform movement needed a new strategy: test-based accountability (Amrein & Berliner, 2002; Ravitch, 2010).

During the 1980s and 1990s, as the nation turned toward standards-based accountability, social promotion also became a more substantive issue in education. Social promotion is the practice of promoting a student to the next grade level even if the student has not achieved academic competencies for the current grade level. Ending this practice became a major priority for both the Clinton administration and state governments, and the proposed solution was to implement end-of-grade testing programs to determine if students had the requisite skills to be promoted to the next grade (Heubert & Hauser, 1999). It was also during these years that the NAEP was transformed from a little-known test to a highly controversial and much debated part of education. In the formative years of NAEP, the politics of the time prevented the release of results at the state level, instead reporting only national and regional data on student performance. By the 1980s, however, this had changed after stakeholders argued for the need for disaggregated state-level data from NAEP results (Bracey, 1995). With this change, state-by-state comparisons were now readily available, as was the ability to compare results from NAEP with results from state-administered standardized tests.

In addition to a focus on standards and test-based accountability, this period was also characterized by the beginning of a shift in accountability from the state level to the federal level (Dwyer, 2004). The 1994 reauthorization of the Elementary and Secondary Education Act

(ESEA), which was titled the Improving America's Schools Act (IASA), required states receiving federal funding to test students at three points between K-12. IASA did not mandate that schools achieve certain levels of proficiency but did ask schools to make yearly progress. However, there was no enforcement of this provision by the U.S. Department of Education. If schools failed to show yearly progress, there were no substantive penalties enforced by the DOE. This changed with the subsequent reauthorization of ESEA as the No Child Left Behind Act.

The No Child Left Behind era. In the years leading up to the passage of NCLB, all but one state established standards and some form of high-stakes testing, effectively paving the way for federal policy (Hush, 2005; Manna, 2006). The testing movement became a national political phenomenon with the passage of NCLB in 2002. NCLB has been described as "the most sweeping plan to shake up public education in a generation" (Dillon, 2004) and "changed the nature of public schooling across the nation" (Ravitch, 2010, p. 15). While standardized testing was already prominent in states across the country (and was often linked to accountability), the passage of NCLB put high-stakes testing and accountability in the national school reform spotlight. Test scores became the primary measure of school quality (Ravitch, 2010).

NCLB is based on the assumption that measuring student progress toward standards for performance will leverage the improvement of student learning. But it also serves motivational and symbolic purposes by establishing targets for reform efforts (Haertel & Herman, 2005). For example, one of the notable and more controversial aspects of the law was the goal of having 100 percent of students across the nation meet or exceed proficiency levels by 2014. And, unlike IASA before it, NCLB had teeth. Schools and districts were now being held accountable by the federal government and penalties for failing to adhere to NCLB's mandates were substantial and, more importantly, enforced. Schools that failed to make adequate yearly progress were subject

to significant sanctions. Penalties included replacing administration and staff, allowing students to move to higher-performing schools, and, in extreme cases, closing down a school altogether.

Common Core State Standards and the Every Student Succeeds Act. Today, the testing debate is largely defined by the Common Core State Standards (Common Core) and the passage in December 2015 of the Every Student Succeeds Act, which reauthorized the Elementary and Secondary Education Act and replaced the No Child Left Behind Act of 2001. The adoption of the Common Core by most states when they were unveiled in 2010 created a need for new assessment systems. Several consortia of states were created to develop these assessments and two continue today: the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers (PARCC), which both have developed assessments that are designed to reflect the standards detailed in the Common Core. Because these tests are aligned with the Common Core, which has been highly controversial, the tests have also been controversial. Proponents have argued that the new tests are innovative, sophisticated, and a dramatic improvement over prior testing systems (Rothman, 2011). Critics, on the other hand, argue that the new testing systems are no better than previous tests, fail to capitalize on technological and research-based developments in math and English, and are costly to implement (Lu, 2014; Rasmussen, 2015).

ESSA, like NCLB before it, requires states to administer annual statewide assessments (in reading and math in grades 3 through 8 and once in high school) that measure students' progress toward standards. But in a significant change from NCLB, the federal role in overseeing accountability is scaled back. Under ESSA, states can establish their own accountability systems, with some general guidelines on potential indicators provided by the U.S. Department of Education (Blad, 2016). The extent to which ESSA will alter the education

landscape of accountability and testing remains to be seen, as scholars have highlighted that the law has many complexities that will need to be worked out in implementation (Orfield, 2016).

The current study focuses on media coverage over the 20-year period from 1995 to 2015 for two reasons. First, my aim is to investigate how the public debate over testing in schools has evolved in recent years. As noted in Chapter One, developments in the last several years indicate a shift in the public and political response to testing. I explore the extent to which this shift is also apparent in media coverage of the issue. Second, as described in the history of testing section above, the prominence of testing within education policy reform gained substantial national attention and was solidified with the passage of NCLB in 2001. It is therefore important to capture how this policy impacted the debate by including several years of data prior to passage of NCLB. By focusing on the last two decades, I am able to both narrow the focus to an exploration of the recent shift and also include years leading up to the passage of NCLB. Twenty years is also a sufficient length of time to identify patterns and trends in media coverage. In order to better understand the complexity of the debate over this period of time, I next review the literature on the politics of testing.

Politics of Testing

Throughout their long history in American schools, tests have been advocated in the service of numerous and diverse reform efforts and education policies. In the antebellum period, for example, advocates argued that tests would shed light on inequality and raise awareness of the need to integrate schools (Reese, 2013). Many decades later, in the 1980s and 1990s, advocates connected testing to the standards-based reform movement, which called for clearly defined, measurable academic standards for what students need to know and be able to do. States initiated and led the standards-based reform movement throughout the 1980s. This effort

was also a prominent education policy proposal advocated for by both the George H. W. Bush and Clinton administrations (Linn, 2008). Today, Common Core assessments and the implementation of ESSA are central issues in educational reform.

Even in the crowded field of major issues and conflicts within education policy, the issue of testing stands out. The titles of some recent publications provide anecdotal evidence of the level of vitriol in the testing debate: *Testing Wars in the Public Schools*; *Kill the Messenger: The War on Standardized Testing*; and *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education*. Tests have been a controversial aspect of public schooling for many years in part because when tests are linked to accountability structures, they often have the power to influence decisions about what is taught in classrooms. Additionally, test scores have frequently been used as a public indicator of system, school, or educator performance.

Testing is Political

"Assessment policies represent a political solution to an educational problem" (McDonnell, 2005).

For many years, testing in schools was seen primarily as a technical issue, with discussion focused on issues of reliability and validity (McDonnell, 1997). In recent decades, however, testing has shifted from a primarily technical issue to a highly political issue involving a broad range of values and beliefs. In her work on testing and accountability in schools, Lorraine McDonnell discussed the political nature of testing and use as a reform effort. According to McDonnell (2005), high-stakes tests can serve seven different purposes: 1) provide information about systems of schooling, 2) guide instructional decisions, 3) create coherent curriculum, 4) certify that students reach certain levels of achievement, 5) motivate students,

teachers, and parents, 6) act as a lever to change instructional approaches or content, and 7) hold schools and educators accountable. The first four of these purposes are relatively straightforward and not particularly controversial. The last three, however, are infused with social and political values and therefore are controversial and have the potential to create tension among education stakeholders.

In addition to these seven substantive applications, McDonnell (2005) also notes that high-stakes testing is appealing to policymakers as a policy instrument of education reform in response to concerns that public schooling is not sufficiently rigorous or accountable. In addition to the appeal for policymakers as a lever for holding schools accountable, tests also serve another accountability purpose: school performance is one measure used by the public to evaluate politicians' performance (and test data serve as a powerful indicator of school performance) (McDonnell, 2005).

Tests as Policy Instruments of Education Reform

According to McDonnell (1994), policy instruments are "mechanisms that translate substantive policy goals into concrete actions" (p. 397). In a series of articles exploring the use of tests as policy instruments, McDonnell (1994, 1997, 2005) notes how high-stakes testing as an instrument of education policy is highly appealing to politicians and policymakers because it is one method for officials at higher levels of government to influence curriculum and classroom practices. And, although many top-down education reform strategies are not particularly effective at altering local practice (McLaughlin, 1987), research suggests that testing is one of the few top-down strategies that actually do change behavior of administrators and teachers in classrooms (Cohen-Vogel & Rutledge, 2009; Firestone, Mayrowetz, & Fairman, 1998; Herman & Golan, 1993).

Anne Schneider and Helen Ingram's (1990) typology of policy tools is helpful in illustrating the ways in which tests have been used as a policy instrument of education reform. The typology includes five broad categories of instruments (authority, incentive, capacity, hortatory, and learning), which are based on underlying behavioral assumptions. Briefly, authority tools assume that people will do as they are told in a hierarchical system; incentive tools assume that people are motivated by a concrete payoff (or sanction); capacity tools assume that people will act if they have the requisite skills or knowledge; hortatory tools assume that people are motivated from within based on beliefs and values; and learning tools assume that in uncertain situations, learning about possible solutions to problems and experimenting with policy approaches leads to changes in behavior (Schneider & Ingram, 1990).

McDonnell (1994) argued that testing is a hortatory policy tool, in the sense that assessment policy is often based on strong appeals to beliefs and values about schooling. Tests have been tied to beliefs about schooling including excellence, high expectations, and ensuring that all students have access to a high quality education. Important for this study is the idea that testing as a policy instrument has a blurred boundary between *objective information* and *value-laden interpretations* (Stone, 1988). McDonnell describes how tests are value-laden policy instruments. Rather than simply being valued as a source of information about student learning, testing policy is often linked to values about how to encourage improvement in schools, what should be taught, and greater transparency and public awareness of how schools and students are performing (McDonnell, 1997). This blurred boundary creates a tension between proponents of tests who argue that tests are an essential aspect of schooling that provide educators with important information about student learning and areas of deficiency and policymakers who

argue that tests can be used to achieve broader reform actions such as assessing teacher performance or determining resource allocations.

However, policies and programs such as NCLB and other federal and state initiatives directly tying test scores to accountability policies such as closing schools or teacher performance evaluations suggest that high-stakes testing is not only a hortatory policy tool but instead is a combination of an authority tool, a hortatory tool, and an incentive tool. A central idea pushed by policymakers is that high-stakes tests will motivate students and educators to improve academic achievement (Nichols, Glass, Berliner, 2006). Often, this is done through test-based incentives, such as offering teachers a monetary bonus or paying students for improved test scores. Evidence of the effectiveness of these types of test-based incentive programs is mixed. For example, Fryer (2010) reviewed a series of randomized controlled trials for programs offering financial incentives for student achievement in urban school districts and found that incentives for inputs (such as reading books or behavior) were sometimes effective but that incentives for outputs (such as course grades or test scores) were much less effective. In a National Research Council report synthesizing incentive programs based on high-stakes testing published in 2011, the authors found that "the evidence related to the effects on achievement of test-based incentives to schools appears to be modest, limited in both size and applicability" (Hout & Elliott, 2011, p. 82). The use of high-stakes testing as a policy instrument for education reform has placed it at the center of the debate over how to improve public schooling. In the following section, I discuss the major arguments in this debate.

Arguments in the Testing Debate

In this section, I first outline the arguments in support of testing and in opposition to testing (see Table 2.1). I then briefly review the role of interest groups and provide an overview of the perspectives on testing within the education research community.

Table 2.1 Arguments in the Testing Debate

For Testing	Against Testing
Establishes high expectations for students and educators	Encourages teaching and staffing to the test
Measures student learning	Narrows curriculum
Provides objective comparable data on performance	Part of a market-based reform agenda
Diagnoses areas of student weakness	Introduces racial and cultural bias
Calls attention to achievement gaps	Encourages cheating
More reliable and valid than alternatives	Misaligned with aims of education

Support for testing. In a historical review of the validity arguments on high-stakes testing in schools, Haertel and Herman (2005) argue that two broad functions of testing have existed since the beginning of testing and accountability in schools. The two functions are 1) to sort or select students (for example, college admissions tests such as the SAT and ACT serve this function) and 2) to improve the quality of education. These broad functions serve as the foundation for the various arguments in favor of testing in schools (see Table 2.1). Furthermore, Hursh (2008) argues that there are three main discourses in the promotion of high-stakes testing: 1) the need to enhance economic and educational productivity, 2) the need to decrease inequality in schooling, and 3) to foster objectivity in measurement.

One of the central arguments in favor of testing is the idea that it is necessary in order to determine if learning is occurring in schools. Haertel and Herman (2005) put it rather succinctly when they wrote, "Students come to school to learn. Tests show which students, in which schools, are meeting learning standards and which are not" (p. 1). Without tests, according to this line of thinking, it is difficult to measure student learning. Related to this reasoning is the argument that tests hold educators accountable for ensuring that all students learn what they are expected to learn and that high-stakes testing helps to establish high expectations for students and educators. Scholars who support testing have noted that the use of tests to measure student learning is a necessary and beneficial aspect of education and that educational accountability is a fundamental right that serves the public interest (Cizek, 2001; Herman, 2008; Ryan, 2008). High-stakes tests can serve as motivation for students to work harder, learn more, and take testing seriously, which in turn can lead to increased achievement (Jacob, 2005).

Other arguments for high-stakes testing center on the idea that tests are an important metric or indicator of educational performance, as well as being relatively objective. High-stakes tests that are standardized and externally enforced also provide easily comparable data on students and schools (McDonnell, 2008). Specifically, test scores provide straightforward, understandable, and comparable data on the performance of schools and students, which has broad appeal among policymakers. Educational leaders and policymakers can use these data to develop and target reform efforts to improve schooling. Test data are also used to compare performance among groups of students by demographic factors, which can highlight achievement gaps between subgroups of students. Increased attention to achievement gaps can thereby prompt targeted reform efforts or resource allocations to address these gaps.

At the student level, test data can be used to diagnose learning difficulties or areas of weakness that can be targeted for additional instruction (Evers & Walberg, 2004). For example, Foorman, Fletcher, and Francis (2004) argue for the importance of assessing early reading skills and show that assessment results, when utilized properly, can help to enhance learning outcomes. Along the same lines, test score data are also useful for parents, postsecondary institutions, and employers as a proxy for level of learning and skill attainment (Phelps, 2003).

Proponents of high-stakes tests also argue that these methods of assessment are more reliable, valid, and objective than other forms of assessment. Some studies suggest that alternatives such as portfolio assessments or performance assessments have poor reliability, are relatively expensive, require extensive teacher and student time, and are subject to potential legal challenges when tied to accountability (Mehrens, 2004; Stecher, 2004). For example, Mehrens (2004) argued that multiple-choice tests are a better measure of students' knowledge and skills than alternative assessment methods.

Additionally, test score data can be used to make better-informed decisions. As Cizek (2005) noted, there is an unavoidable need for educators and policymakers to make decisions and these decisions should be based on the best available information. If high stakes tests provide the most reliable measure of student performance, then these tests should be used to inform decisions. Proponents have also argued that in spite of claims that tests are the sole measure used to make high-stakes decisions, they are rarely, if ever, the only piece of information that is used (Cizek, 2001). At the core of this argument is the idea that until better, higher quality alternative assessments are developed, standardized tests are the best available option for educators and policymakers.

Critiques of testing. In addition to the arguments in favor of high-stakes testing, several major criticisms have been articulated. One of these criticisms is that a focus on testing alters instruction in negative ways. First, because of the high stakes attached to the tests, educators may feel pressure to "teach to the test" (Au, 2007; Jennings & Rentner, 2006; Madaus, 1988). Focusing on the content covered in the tests can lead to important but untested concepts being deemphasized and, depending on the nature of the test, may result in a focus on memorization at the expense of critical thinking skills. Critics have argued that high-stakes testing practices are misaligned to the overarching aims of education, such as developing critical thinking (Siegel, 2004). Similarly, critics argue that testing promotes a narrow curricular focus in schools because educators may reduce instruction in untested subject areas such as art, health, or social studies (Dillon, 2006a; Herman & Golan, 1990; Shepard & Dougherty, 1991).

Additionally, school leaders may feel pressure to "staff to the test" by reallocating teacher resources in an effort to manipulate school performance (Cohen-Vogel, 2011; Grissom, Kalogrides, & Loeb, 2015). This is because pressures to increase student performance apply disproportionately to tested grades and subjects (i.e. high-stakes classrooms). Scholars have found that school leaders respond to these pressures by assigning teachers who produce greater student achievement gains to high-stakes classrooms. This strategic maneuvering can be problematic when less effective teachers are assigned to non-tested subjects or grades (Grissom, Kalogrides, & Loeb, 2015).

Another major critique centers on the impact of high-stakes testing on students. In some research studies, testing is associated with increased failure rates, lower graduation rates, and higher dropout rates, particularly for minority groups (Heilig & Darling-Hammond, 2008), students from low-income households (Herman and Baker, 2009), students with special needs

(Katsiyannis, Zhang, Ryan & Jones, 2007), and students with limited proficiency in English (Abedi, 2005; Fine & Jaffe-Walter, 2007). Critics also argue that testing decreases student self-esteem (Meisels, 2000) and contributes to negative attitudes toward content covered on tests (Lattimore, 2001). A related issue that critics have pointed to is that testing exacerbates negative stereotypes about the intelligence and academic ability of minority students (Phelps, 2003), who may also suffer from stereotype threat, a phenomenon in which students who belong to a specific group that is associated with a negative stereotype are impacted in their performance on high-stakes tests when the stereotype is brought to their attention (Steele, 1997).

The high-stakes associated with testing may also contribute to bad behavior among educators and students, specifically higher rates of cheating. Under pressure to improve student achievement, teachers and administrators may change student test scores in an attempt to preserve their reputations and jobs. Recent high profile scandals such as the systematic cheating among educators in the Atlanta public schools have drawn considerable attention to this negative consequence of high-stakes testing (Gabriel, 2010; Perry, Judd, & Pell, 2012). In an analysis of the prevalence and predictors of teacher cheating, Jacob and Levitt (2003) found that teacher cheating rates among the lowest performing classrooms in Chicago Public Schools increased after test scores were linked to accountability policies. Recent research even suggests that teacher cheating through test score manipulation may lead to lower subsequent student achievement (Apperson, Bueno, & Sass, 2016).

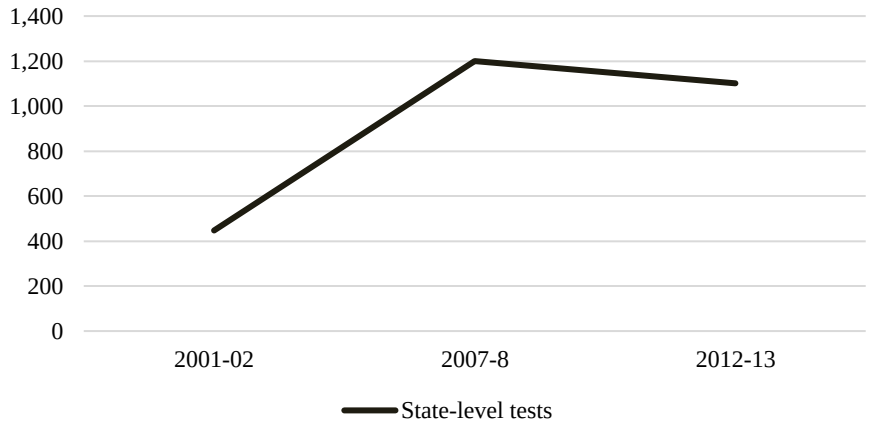
There is also a political dimension to the arguments for and against the use of high-stakes testing in schools. In a discussion of the persistence of high-stakes testing in schools, Moses and Nanna (2007) note, "The use of high stakes testing is related to political ideology and the exercise of political will and power instead of being driven primarily by the best interests of

students" (p. 63). Proponents of testing claim that standardized tests offer a more objective and egalitarian measure of student performance than other possible measures, in addition to providing important information on whether or not different student subgroups are receiving quality schooling. Therefore, proponents advocate for testing as a means of creating more equitable schooling across student populations. In contrast, some critics have argued that testing is part of a larger neoliberal agenda promoting choice, competition, and accountability in public education, with little to no specific regard for creating equity in public schooling (Croft, Roberts, & Stenhouse, 2016). Many of these arguments for and against testing have been utilized by interest groups in support of their positions on testing and test-based accountability. These groups are discussed in the following section.

Interest Groups. According to McDonnell (2005, 2008), the interest group environment for high-stakes testing and accountability has three main characteristics. First, the environment features numerous and diverse groups. Although these groups generally support the same broad goal of improving schools and student learning, they differ in the specific strategies and reforms they support and the role of testing in achieving this goal. Second, these groups do not align along partisan lines and, further, organizations with similar purposes often have different positions on testing. For example, despite some evidence that high-stakes testing may not be helping disadvantaged groups of students, in early 2015, a coalition of 27 civil rights groups, including the NAACP, the National Urban League, and the League of United Latin American Citizens (LULAC) signed a statement calling on Congress to maintain the testing requirement in the reauthorization of ESEA (Ehrenfreund, 2015). Additionally, although the national LULAC signed the statement, local chapters expressed disagreement, an indication that even within organizations, positions vary (Michaels, 2015). The third characteristic is that the interest group

environment features a unique relationship between "institutionalized elements of test-based accountability and the interests that benefit from them" (McDonnell, 2008, p. 55). The result of this relationship is that different interests regard the debate over high-stakes testing from different perspectives. The testing industry, for example, benefits from policies related to testing students on a regular basis. As shown in Figure 2.1, U.S. spending on state-level testing increased from $447 million to $1.1 billion in the decade after the passage of NCLB (Cavanagh, 2015). Politicians, in contrast, often support testing and accountability as an electoral strategy.

Figure 2.1 U.S. Spending on State Assessments (in millions)



Proponents of high-stakes testing include organizations that believe it is an effective reform strategy to improve academic achievement, such as Achieve, Education Consumers Clearinghouse, the Education Trust, and the national Business Roundtable. The testing industry is a major supporter primarily for financial reasons. Additionally, some civil rights groups such as the NAACP and the National Council of La Raza support high-stakes testing as an effective way to call attention to and address the achievement gap (Leadership Conference, 2015; McDonnell, 2005, 2008; Phelps, 2003). Critics of high-stakes testing include teachers unions, FairTest, United Opt Out, and various state and local organizations. National teachers unions (the National Education Association and the American Federation of Teachers) were particularly

vocal about NCLB and have argued that important accountability decisions should not be based solely on tests.

Research on High-Stakes Testing

Perhaps not surprisingly, there is little consensus about high-stakes testing within the education research community. Some studies highlight the positive aspects of high-stakes testing and accountability policies. For example, an edited volume on the benefits of testing from 2004 includes a list of over 350 studies that indicate achievement benefits of testing (Phelps, 2005). Others indicate that testing and accountability pressures are detrimental and that the negative consequences of high-stakes testing outweigh the potential benefits. While a complete review of the research literature on testing is beyond the scope of this study, a few examples will help to highlight the lack of consensus.

Research indicates that test-based accountability pressure can increase student performance (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Jacob & Lefgren, 2004; Reback, Rockoff, & Schwartz, 2014; Wong, Cook, & Steiner, 2009). Using an index to rate the strength of accountability in all 50 states based on high-stakes testing, Carnoy and Loeb (2002) investigated whether the accountability index was associated with several measures of student performance: scores on the NAEP, rates of grade retention, and progression rates to 12th grade. The researchers found that gains in performance on the NAEP were larger in states with stronger accountability systems and found no apparent negative impact of accountability on student retention or progression rates.

In another study of the impact of test-based accountability on student performance, Jacob (2005) used an interrupted time-series design and longitudinal student-level data to compare the achievement of Chicago public school students before and after the introduction of a high-stakes

testing policy. He found that the testing policy sharply increased student achievement in math and reading. However, the analysis also revealed an increase in the proportion of students in special education and an increase in student retention after the policy was implemented. In an analysis of the impact of test-based accountability pressures under NCLB, Reback, Rockoff, and Schwartz (2014) used data from the Early Childhood Longitudinal Study and found that accountability pressures had either a positive (or in some cases neutral) effect on student achievement and reported no adverse impact on students' reported anxiety about testing or enjoyment of learning. The authors also reported no heterogeneity of effects across student subgroups. Taken together, these studies suggest a positive impact of test-based accountability on student performance.

However, other research suggests that some students may benefit more than others from increased accountability pressure (Booher-Jennings, 2005; Ladd & Lauen, 2010; Lauen & Gaddis, 2012). Specifically, studies indicate that students near cut scores for proficiency levels (also known as "bubble kids") are targeted for additional support by educators, a phenomenon known as "educational triage" (Booher-Jennings, 2005; Brown & Clift, 2010; Krieg, 2008; Lauen & Gaddis, 2012). Educational triage is potentially problematic if it means that not all students are receiving equal benefits from accountability pressure. Other research suggests that pressure from high-stakes testing does not influence academic achievement (Nichols, Glass, & Berliner, 2006). Scholars have also analyzed high school exit exams and GED rates and found that more challenging exit exams were associated with lower completion rates, providing further evidence that accountability pressures may not be leading to academic improvements (Warren, Jenkins, & Kulick, 2006).

Overall, there is ongoing ambiguity within the research community about the effects of high-stakes testing and test-based accountability on student achievement. This brief review provides further evidence of the complexity and ambiguity of the testing debate. If the research on the impact of high-stakes testing was unambiguously and strongly positive or negative, this issue might not be particularly controversial. But with little consensus on the effects of testing, the issue remains both controversial and open to various interpretations and understandings.

Scholars have sought to explain why test-based accountability, despite mixed findings of its usefulness and some potentially negative consequences, continues to be a popular school reform policy. Moses & Nanna (2007) propose four factors that contribute to the continued popularity of high-stakes testing (beyond the primary factor - the need to assess student learning). The four factors are 1) administrative utility, 2) profit motives, 3) political ideology, and 4) the testing culture. Tests are viewed as administratively appealing for being an efficient, cost-effective, straightforward solution to the challenge of processes high volumes of information to make decisions. A classic example is the use of SAT or ACT scores as an efficient way to screen thousands of college admissions applications. Testing has also become a huge industry, which may contribute to its continued popularity. As discussed earlier, test-based accountability is closely linked to political ideology and tests have been a popular policy tool with politicians from both sides of the aisle. Finally, testing remains popular in schools because testing in general has become an ingrained part of our culture. Tests also appeal to a culture enamored of numerical measures of success. In addition to these four reasons, other reasons why high-stakes testing remains a popular education reform strategy are that testing is an inexpensive reform relative to other possible options for changing instruction or curricula; tests can be

externally mandated; tests (and revisions to tests) can be implemented relatively quickly; and results are visible, thereby increasing transparency (Linn, 2000).

Given that research on high-stakes testing is mixed, some scholars in the education research community argue that testing has the potential to serve as an important aspect of educational measurement and reform efforts but that the tests themselves must first be improved. Additionally, scholars argue that careful consideration must be given to the appropriate uses of test score data as well as what exactly schools should be held accountable for (Darling-Hammond, 1991; Dwyer, 2004; McDonnell, 2008). Regardless of their stance, most scholars are in agreement that the entrenched culture of testing in American schools is not likely to disappear in the near future, particularly because testing is frequently advocated as part of more comprehensive reform strategies (Haertel & Herman, 2005; McDonnell, 2008; Moses & Nanna, 2007; Nichols, Glass, & Berliner, 2006, Ryan & Shepard, 2008).

The arguments supporting and critiquing testing, as well as the mixed findings about the effects of testing and test-based accountability policies, indicate that this is a complex, multidimensional issue. Additionally, interest group influence, policymaker and stakeholder opposition or support, and the long history and institutionalized culture of testing in American schools all are important factors in analyzing the dimensions and frames of the issue of testing. Next, I turn to public perceptions of testing.

Public Opinion and Testing

An important aspect of the politics of testing is public opinion about the issue. Public opinion can play an important role in policy making (McDonnell, 2005; Page & Shapiro, 1983). Generally, satisfaction and confidence in public institutions has been on a downward trend for the past 50 years, and specifically in education for the past 30 years (Jacobsen, 2009). Although

satisfaction in education generally has been declining, there has been broad support for many years for the general concept of testing as a method of measuring academic progress, raising standards, and evaluating schools and teachers (Behuniak, 2003; Cizek, 2005; Johnson, 2013; McDonnell, 2005; Phelps, 1998). An analysis of test-related sections on public opinion polls and surveys over almost 30 years (1970-1997) revealed a fairly consistent pattern: during that time period, the majority of respondents were in favor of more high-stakes testing or higher stakes for current testing (Phelps, 1998). During the 1970s, a large majority of the public supported the use of tests as objective measures of student achievement. Testing continued to receive high levels of public support (according to polls) throughout the 1980s and 1990s. In these decades, the idea of comparing across schools and districts gained prominence, as did international comparisons. Brookhart (2013) argued that testing has appeal to the public in part because test scores function as an indicator in the international competition for economic superiority. Public support for the general idea of testing continued through the 2000s, but there was less support for specifics related to NCLB, such as using test scores in evaluating teacher or administrator performance (Brookhart, 2013).

However, even with overall levels of support for testing in schools remaining high, there are signs of a relatively small but growing public backlash. Surveys of the public and parents of public school students in the last decade suggest that there are more nuanced perspectives about the role of testing (Public Agenda, 2006). For example, in focus groups with parents in a sample of cities across the country, parents expressed that testing was important but that it was overly emphasized and that it should be one of multiple measures for evaluating schools, teachers, and students (Johnson, 2013).

This testing backlash has gained more attention as suburban, middle class parents have voiced concern. The main arguments put forth by these parent groups are that test preparation stifles classroom creativity, tests are time consuming, testing creates undue stress for students, and testing disadvantages some students who are not good test takers (McDonnell, 2005). In addition to indications of a shift in perceptions among survey respondents, there is also a more direct and growing movement among some families dissatisfied with testing in schools – the opt out movement.

The Opt Out movement. Unlike more general public perceptions of testing in schools, which has been studied for many years, scholars are only just now beginning to study the phenomenon of the opt out movement and the motivations of parents and students who have engaged in this form of activism. My search of the literature uncovered only one study of the motivations behind the opt out movement, even though the movement has received substantial media coverage in the past three years. In a survey of over 1,600 opt out parents across 47 states, Pizmony-Levy and Green Saraisky (2016) found that the opt out movement is not only about opposition to tests but is also more generally critical of reform efforts that focus on test-based accountability. The researchers also found that the movement is dominated by opt out activists who are white, married, highly educated, and wealthier than the average American.

It remains unclear how many families opt out of tests. FairTest: The National Center for Fair and Open Testing, which is a strong voice in the anti-testing movement, estimates that approximately 670,000 students opted out of public school testing in 2015, but this number is based on news reports and surveys by local activists (FairTest, 2016). This is also a very small portion of the approximately 50 million students in public elementary and secondary schools

(U.S. Department of Education, 2015b), and surveys suggest that the public generally, and parents specifically, still overwhelmingly support annual testing (Felton, 2015b).

The current opt-out movement has received substantial media coverage and appears to be a relatively widespread and growing backlash to testing policies. However, grassroots organizing by parents in response to testing requirements is not completely new. More modest efforts were reported in several states, including New York (Zernike, 2001) and Massachusetts (Greenberger, 2000), years before the current movement began in 2014. What is clear, however, is that the current opt out movement is gaining momentum and growing in states across the nation (Strauss, 2016; Schweig, 2016; Ujifusa, 2016). In New York State, for example, the growing opt out movement has had several repercussions. The movement helped to prompt the governor to appoint a task force to review the state's implementation of Common Core State Standards. Additionally, it contributed to changes in leadership in the state's board of regents and led to a public relations campaign across the state by the education commissioner in response to the movement (Burnette, 2016).

There is also some indication that policymakers at the national level are paying attention to the opt out movement (Felton, 2015a). Specifically, the reauthorized Elementary and Secondary Education Act, which reduces the federal role in testing and accountability and calls on states to develop plans for addressing student opt outs, suggests that policymakers are, in fact, attuned to and responding to the growing backlash against testing.

Framing

In this section, I define framing, describe key concepts and aspects of framing, and describe the conceptual framework that guides this study. Key concepts in framing include the

multidimensionality of issues, the distinction between emphasis and equivalence framing, and characteristics of issue-specific framing.

In the last several decades, scholars have given substantial attention to the concept of framing, but there is little consistency in definitions of the concept (Cacciatore, Scheufele, & Iyengar, 2016). To illustrate, here are a few definitions from the framing literature:

- Framing refers to subtle alterations in the statement or presentation of problems (Iyengar, 1991, p.11).
- Frames are organizing principles that are socially shared and persistent over time, that work symbolically to meaningfully structure the social world (Reese, 2001, p. 11).
- Framing refers to the process by which people develop a particular conceptualization of an issue or reorient their thinking about an issue (Chong & Druckman, 2007, p. 104).

For the purposes of the current study, I adopt Entman's (1993) definition of framing: "To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described" (p. 52). In conceptualizing framing as a theory of media effects, scholars have also divided framing into two linked categories: media frames and individual (or audience) frames (Scheufele, 1999). This study is concerned with media frames, which involve the presentation of information or news. The framing of information by the media can systematically affect how individuals understand problems or events.

Issue-Specific Emphasis Framing

Framing can occur in various ways. In their essay on the future of framing effects research, Cacciatore, Scheufele and Iyengar (2016) discuss the importance of clarifying the

conceptualization of framing and distinguish between emphasis framing and equivalence framing. Emphasis framing involves "manipulating the content of a communication" while equivalence framing involves "manipulating the presentation of logically equivalent information" (p. 8). The current study focuses on emphasis framing by seeking to understand *what* information audiences receive about high-stakes testing through media coverage of the issue (emphasis framing also aligns with Entman's (1993) definition of framing, which highlights the concepts of *selection* and *salience*).

In addition to distinguishing between emphasis and equivalence framing, scholars have also described assumptions and characteristics of issue-specific framing. Two central assumptions about issues that relate to framing are that 1) policy issues are multidimensional and 2) issue definitions can change or evolve. Issue dimensions are the broad characteristics of the issue (also referred to as attributes) that can be arranged (highlighted or suppressed) by policy actors to (re)define the issue. How issues are understood along various dimensions affects how they are perceived by the public and policymakers and impacts whether they reach the policy agenda (Baumgartner & Mahoney, 2008; Jochim & Jones, 2013; Nowlin, 2016; Rochefort & Cobb, 1993).

Drawing from the work of Baumgartner, De Boef, and Boydstun (2008), I conceptualize framing as including arguments and issue dimensions. Arguments are the smallest unit of framing and include the specific aspects of the issue or debate. For example, in the testing debate, one "argument" would be coverage of the cheating scandals in which educators changed test scores in response to pressure to improve student performance. Issue dimensions are broader, more generalized categories of the debate. Arguments can be clustered into these dimensions. For example, news coverage about school grading systems or teacher evaluations

based on test scores might be categorized under a dimension called performance-based accountability. In my study, I refer to arguments as frame elements.

The definition of an issue can change over time as new dimensions of an issue are selected and highlighted over others. In a multidimensional issue space, policy actors can arrange or emphasize frame elements in different ways, leading to different definitions and new interpretations of the issue over time (Baumgartner & Mahoney, 2008).[2] Rochefort and Cobb (1993) noted, "Viewed over a sufficient span of time, this evolutionary pattern of issue transformation is perhaps more the rule than the exception within our dynamic political environment" (p. 56).

In addition to policy actors, media also engage in this process of framing (Gamson, 1992; Iyengar, 1991; McCombs & Shaw, 1972; Tewksbury & Scheufele, 2009). An excellent illustration of the influence of the media on policy issues and how media framing can result in new understandings of issues is Baumgartner and Jones' (2009) work on the dynamics of media attention. The authors discuss how media coverage rarely focuses on multiple aspects or dimensions of a given policy issue. Instead, issue coverage is often limited to only one or a few dimensions of a complex issue at any given time. The media can shift the focus to different dimensions of the debate, and when this happens, it results in new understandings of the issue. The authors state:

> When attention shifts to another dimension of the same policy, the tone of the coverage can be reversed dramatically. Even in the absence of new findings, new scientific

[2] The dimensions dominating any policy debate are partly determined by these efforts by policy actors to manipulate issue dimensions, but are also determined by external factors such as random events or crises and new information. This is discussed in detail in Jones and Baumgartner, 2005.

> evidence, or new arguments, the nature of debate surrounding an issue can shift dramatically if only it shifts in focus (Baumgartner & Jones, 2009. p. 109).

This notion of shifting attention within media coverage of an issue is central to longitudinal studies of how understandings and definitions of issues change over time. Issue-specific framing, then, can be understood as the process of manipulating dimensions (through addition, subtraction, or shifts in emphasis of frame elements) within the multidimensional issue space in order to define an issue in a specific way.

As Baumgartner, De Boef, and Boydstun (2008) note, "Framing may occur through use of a single argument, a cluster of arguments with a single dimension, or even a cluster of arguments across different dimensions" (p. 107). In other words, one frame element can itself constitute a frame or several frame elements within one dimension or across several dimensions can cluster together to form a frame. The current study is concerned with investigating issue-specific framing, also referred to as second-level agenda setting (McCombs & Ghanem, 2001; Shaw & McCombs, 1977). There are numerous examples of research on issue-specific media framing, including works on poverty (Rose & Baumgartner, 2013), the death penalty (Baumgartner, DeBoef, & Boydstun, 2008), same-sex marriage (Wiggins, 2001), and the war on terror (Boydstun & Glazier, 2013).

In their work on the framing of poverty, Rose and Baumgartner (2013) analyzed all stories on poverty in the *New York Times Index* from 1960 to 2007 to measure media attention to the issue across time. The authors grouped the many different arguments into five distinct frames: three frames with a positive or "generous" tone toward poverty (economic/physical barriers, misery/neglect, and social disorder) and two with a negative or "stingy" tone (laziness/dysfunction and cheating). The early period of media coverage was dominated by the

generous frames but over time attention shifted dramatically to a predominance of the laziness frame in recent decades.

Similarly, Baumgartner, De Boef, and Boydstun (2008) analyzed an extended time period of media attention (again using the *Times Index*) to track the shifting terms of the debate in media coverage of the death penalty. The authors developed a novel statistical approach called evolutionary factor analysis to identify patterns of arguments that cluster together to dominate the debate in different periods of time and document the rise of a powerful new frame: innocence. Studying public debate on same-sex marriage, Wiggins (2001) analyzed over 100 letters to the editor reacting to a news article on a lesbian commitment ceremony in a local newspaper in the South. Letters were categorized as either supporting or opposing publication of the story and religious terms and civil rights terms were identified to assess the relationship between the letter writer's position and the use of religious versus civil rights terminology. No significant relationship was found.

Boydstun and Glazier (2013) take a slightly different approach in their work on media framing by identifying not only issue-specific frames but also broader, generalizable frames (gain-based and loss-based frames) that can be applied to multiple issues. This "two-tiered" framing approach was applied to coverage of the war on terror in the *New York Times* and *Wall Street Journal* between 2001 and 2006. By including the generalizable frames in the analysis, the authors were able to identify broad trends over time in addition to issue-specific frame elements. The authors found that media framing following the September 11 attack initially focused on loss-based frames (e.g. "Terrorists hate our freedom.") but shifted to gain-based frames (e.g. "More democracies in the world gives us more allies.") over time, suggesting that

public willingness to support war in Afghanistan and Iraq may have been influenced by media framing of the issue.

In the current study, I identify frame elements in media coverage of the issue of testing in schools and explore how they contribute to shifts in the understanding of the issue. In the following sections, I discuss how scholars have approached the process of measuring frames and I outline the conceptual framework used to explore the process of issue transformation.

Measuring Frames

One of the most notable characteristics of frames is that they are difficult to measure. For many years, scholars of framing have wrestled with the methodological challenges regarding reliability and validity of measuring frames (Matthes, 2009; Matthes & Kohring, 2008; Miller, 1997; Tankard, 2001). Matthes and Kohring (2008), for example, describe five dominant content analytic methods employed in studies of media framing and argue that all five suffer from either problems with validity or problems with reliability. Instead, they put forth a method of frame analysis that does not attempt to code entire frames but instead (drawing on Entman's definition of framing) focuses the analytic lens on frame elements, which they argue are easier to identify and label. As noted earlier, frame elements can be understood as specific arguments in a policy debate. Combinations of frame elements make up a frame. Frames, then, can be identified as clusters of these easier-to-measure frame elements. This approach to measuring frames is conceptually similar to the approach of Nowlin (2016) and Baumgartner, De Boef, and Boydstun (2008). Nowlin (2016) applies automated content analysis methods to identify the dimensions that collectively define the issue of management of used nuclear fuel. Baumgartner, De Boef, and Boydstun (2008) use an approach called evolutionary factor analysis to identify

patterns in argument use and explain how clusters of arguments can develop into powerful frames.

In addition to identifying frame elements and exploring how these elements cluster together to create frames, I also measure the power of frames by identifying three characteristics of frames: salience, resonance, and persistence. Together, these characteristics can help to clarify the capacity of frames to influence. Higher levels of these measures of frames equals more powerful frames. In their work on the death penalty debate, Baumgartner et al. (2008) use these three measures to identify the power of different frames over time and offer the following definitions: "First is *salience*, or how often a related set of arguments is used. Second is *resonance*, or how many individual arguments cluster together to constitute the frame. Third is *persistence*, or how long a frame lasts" (p. 138). I use these characteristics of frames in the current study.

Salience is a measure of the prominence of a frame and its associated frame elements. I measure salience by identifying the proportions of frame elements in the collection of news articles (proportionality is estimated through topic modeling, which is described in Chapter Three). Resonance is a measure of the extent to which frame elements *work together* to create a frame and define the debate. The composition of frames can evolve over time as elements couple or decouple from frames. The higher number of frame elements that cluster together, the greater the resonance of the frame. Persistence is a measure of how long a frame lasts. Additional details about these characteristics are provided below in the discussion of the conceptual framework.

Another important concept from the framing literature is cultural resonance (Entman, 2004). Cultural resonance, according to Gamson and Modigliani (1989), is one of the

determinants of successful frames in media discourse. Certain interpretations in the media have an advantage because the ideas and language of these interpretations resonate with larger cultural themes, values, or experiences (Miller & Riechert, 2001). Cultural resonance increases the appeal and potential for influence of a frame by making it appear familiar and innate (Gamson & Modigliani, 1989). Snow and Benford (1988) make a similar point in discussing the "narrative fidelity" of a frame. Some frames "resonate with cultural narrations, that is, with the stories, myths, and folktales that are part and parcel of one's cultural heritage" (p. 210). This notion of cultural resonance may be at play with the framing of testing. If the larger cultural theme of testing is negative, media frames that resonate with this negative theme are likely to have a greater impact on public perceptions. However, if the larger theme is positive, then positive framings of testing in the media are likely to have a larger impact on perceptions.

Conceptual Framework

For the purposes of the current study, I developed a conceptual framework of framing that has five key features based on the literature reviewed above (see Figures 2.2, 2.3, and 2.4). First, frames are composed of frame elements, which are identified in the topic modeling process (see Figure 2.2). These frame elements can include news articles with arguments in support of or critiquing testing as well as coverage of specific aspects of the issue, such as testing scandals. Frame elements can also be composed of topics that are indirectly related to aspects of testing, such as the Common Core State Standards. Second, frames are dynamic and evolutionary. That is, frames can evolve over time by the addition or subtraction of frame elements. Similarly, frame elements can couple and decouple to different frames over time. Figure 2.3 illustrates this concept. Frame Element A contributes to Frame 1 in Time 1 and also to Frame 2 in Time 2. Third, frame elements are measured by their salience, which is the amount of attention (media

coverage) the element receives over time. For the current study, level of attention is measured by proportionality in the topic modeling process (topic proportionality is explained in Chapter Three). Topics (frame elements) with higher relative proportionality in the collection of news articles are more salient than those with lower proportions and this proportionality can change over time. For example, a frame element describing the opt out movement would have very low levels of media coverage for much of the 20-year period of the study but would have a relatively higher level of attention (and therefore higher salience) in the last several years.

Fourth, the number of frame elements that make up a frame is a measure of frame resonance. Resonance, like salience, can change over time. There may be periods when a frame is comprised of more or fewer frame elements. In Figure 2.4, this is represented in Frame 2, which is comprised of four frame elements (A through D) for several years but only three elements in the middle period and near the end of the prominence of this frame. Fifth, the length of time that frame elements cluster together to create a frame is a measure of frame persistence. The length of the horizontal bars in Figure 2.4 represents persistence. For example, Frame 2 is a more persistent frame than Frame 1 or Frame 3. Together, salience, resonance, and persistence are a measure of the power of a frame.

Figure 2.2. Identifying Frame Elements and Frames from the Collection of Newspaper Articles Using Topic Modeling

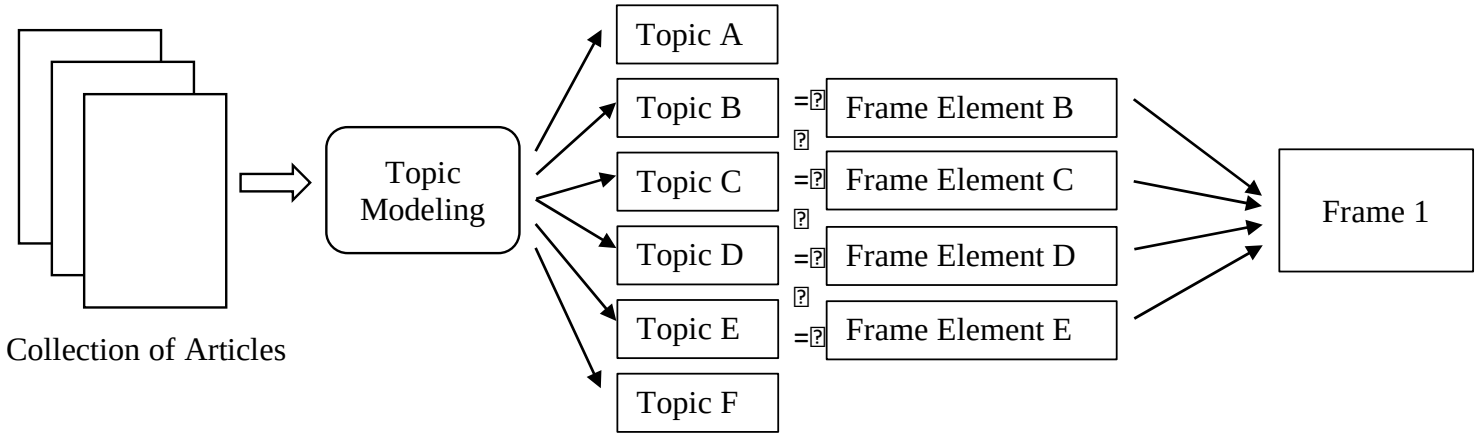


Figure 2.3. Frame Elements Couple and Decouple to Different Frames Over Time

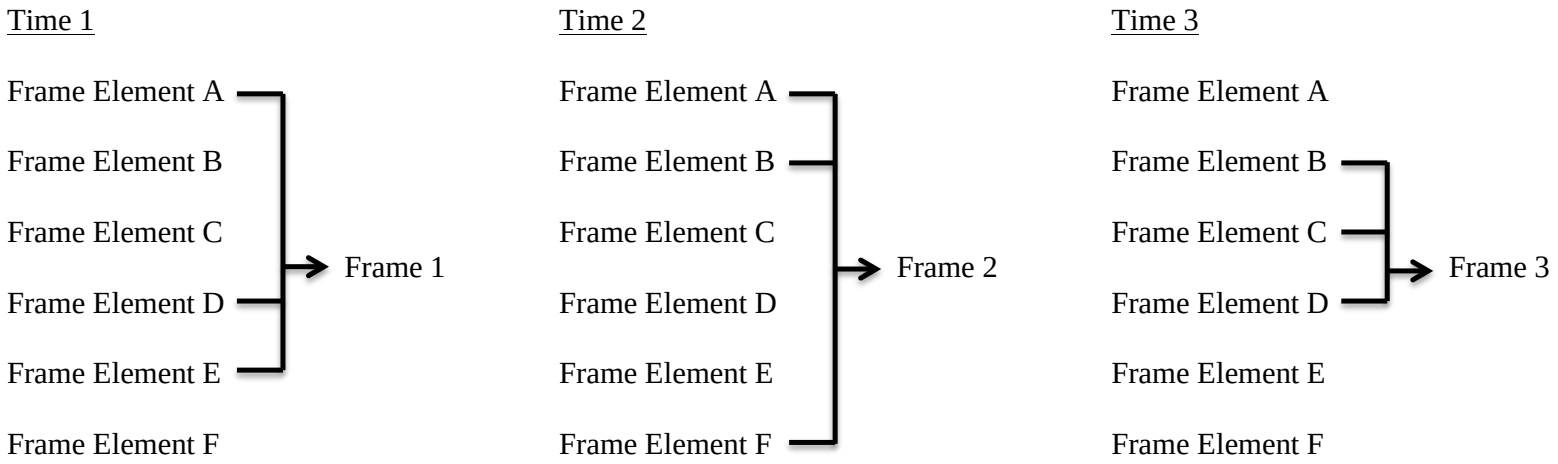
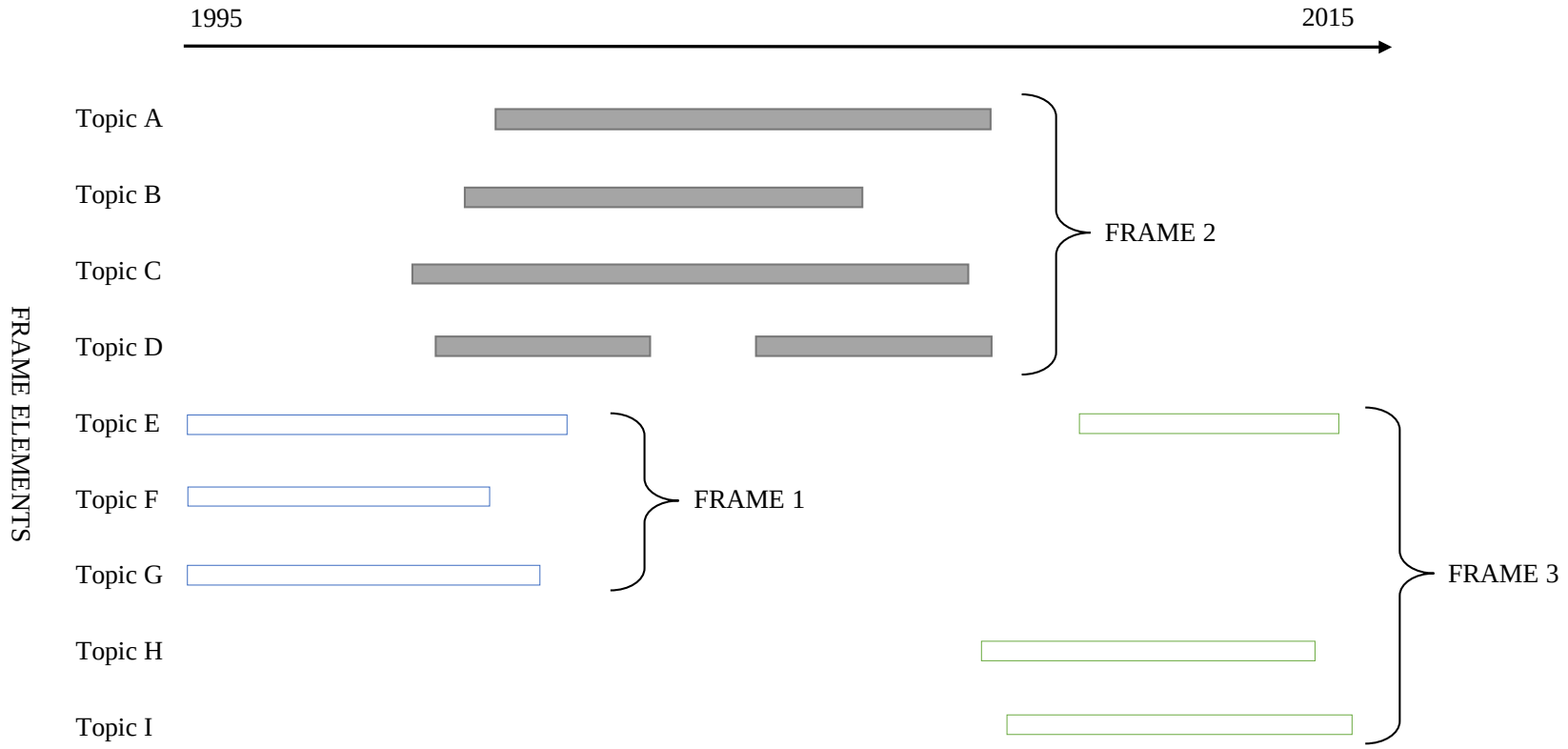


Figure 2.4. Conceptual Model for Frame Salience, Resonance, and Persistence



Note: The output of the topic model produces Topics A through I. The topics are operationalized as frame elements that combine to make frames (also sometimes referred to as issue dimensions). The horizontal bars represent periods of time when the frame elements have relatively higher proportionality (saliency) in the collection of documents (as identified through the topic modeling process). Combinations of frame elements make up frames.

This conceptual framework is used to guide the analysis, including identifying frame elements and measuring their salience across time, measuring the resonance and persistence of frames, tracking the evolution of frames, and, ultimately, answering the research questions outlined above. In Chapter Three, I describe the specific methods used to identify the framing of the issue of testing in media coverage and whether and how the framing changes over time. In the next section, I review the literature on the role of the media in public policy and public perceptions.

The Role of Media in Shaping Public Policy and Public Perceptions

Media and Public Perceptions

The influence of media in politics (e.g. elections and citizens' political behavior) is well established. For example, studies have established strong relationships between media and public perceptions and priorities (Behr & Iyengar, 1985; Dearing & Rogers, 1996; Ghanem, 1996; Iyengar & Kinder, 1987; McCombs, 2004; McCombs & Shaw, 1972; McCombs & Shaw, 1993; Miller, 1997). Previous research has shown that both breaking news events (hard news) and human interest or background news stories (soft news) carry implicit policy messages or cues and that audiences are prompted to consider the policy implications (Boydstun, 2013; Price, Tewksbury, & Powers, 1997; Prior, 2003). In an article describing press coverage of government arts funding, DiMaggio, Nag, and Blei (2013) highlighted five ways in which media can influence public perceptions (both individual and collective perceptions). The five ways are by 1) strengthening existing views (priming), 2) developing new perceptions, 3) integrating with broader beliefs, 4) indirectly influencing through re-telling in social interactions, and 5) serving as a proxy for what opinion leaders view as important (DiMaggio, Nag, & Blei, 2013).

Media and Public Policy

The role of media in the policymaking process is more complex than the media's influence on public perceptions (Voltmer & Koch-Baumgarten, 2010). Although there is a literature establishing the existence of a relationship between the media and the public policy process, the exact nature of the relationship has been the subject of a longstanding debate that essentially boils down to *who is leading whom?* Identifying a causal relationship has been problematic. Three theories have dominated the debate: 1) the influence theory, in which the media influence what politicians think, 2) the agenda setting theory, in which the media influence the policy agenda, and 3) the indexing theory, in which politicians influence media coverage (Jones & Wolfe, 2010). The influence of the media on agenda setting and other specific processes of political decision making has received relatively little attention from scholars (Voltmer & Koch-Baumgarten, 2010; Wolfe, Jones, & Baumgartner, 2013). Due to the relative lack of empirical studies, it is difficult to determine the extent to which the media influence the political agenda and under what specific circumstances the media are able to increase attention to issues (Walgrave & Van Aelst, 2006).

In their review of 19 studies of the media influence on policy making, Walgrave and Van Aelst (2006) report that most of the studies found a strong association but others found almost none at all. The authors reason that the contradictory research findings are due to the fact that the media's influence on the policy process depends on numerous moderating factors, a view shared by other scholars (Baumgartner & Jones, 2009; Voltmer & Koch-Baumgarten, 2010; Walgrave & Van Aelst, 2006). Responding to the contradictory findings, scholars have suggested that it may not be particularly important whether the media are agenda setters or indexers. Instead, drawing on work on information processing, allocation of attention, and

complex systems (Baumgartner & Jones, 2005), it is more likely that media attention and government attention *affect each other* in nonlinear, context-specific ways (Baumgartner & Jones, 2009; Jones & Wolfe, 2010). One of the contextual conditions that moderates the degree to which media influence the process of political decision making is the policy subsystem of the issue. Policy subsystems are complex structures of policy actors, procedures, and policy instruments.

In sum, although it remains unclear exactly how the media influence policymaking processes, media attention is an important resource that influences both politics and society (Boydstun, 2013; Iyengar, 1997). Stated succinctly, "The media agenda is simultaneously an input and an output of the political system" (Wolfe, Jones, & Baumgartner, 2013, p. 186).

Media and Education

General consensus among the scholarly community is that when the media report on education, the focus is frequently negative (Berliner & Biddle, 1999; Camara & Shaw, 2012; Phelps, 2003). For example, in an article arguing for the need for more systematic research on the impact of the media on public perceptions of education, Opfer (2007) notes that the majority of the extant literature on the topic is based on an underlying belief that the media contribute to a negative perception of public education. In an article on the impact of the media on education policy, Anderson (2007) reviewed how the "political spectacle" was constructed in several instances in education. The political spectacle is the notion that the media use "coercion, propaganda, and the portrayal of issues in terms that entertain, distort, and shock to extract a public response" (Edelman, 1988, p.7). Anderson (2007) argued that the popular perception that American schools are in crisis is exacerbated by the tendency of the media to focus on controversial or negative aspects of schooling, such as instances of school violence.

In terms of overall coverage, however, the media cover education issues much less frequently than many other topic areas. In a study of national news coverage, West, Whitehurst, and Dionne (2009) determined that less than 1.5 percent of national news coverage during the first nine months of 2009 was reporting on education (and the percentages were even lower for 2007 and 2008). Within these low percentages, education policy issues were particularly rare (more common were stories about education budgets and school crime) (West, Whitehurst, & Dionne, 2009). Because education is not a topic that receives substantial attention from national news outlets, this study includes an analysis of both a national newspaper (*The New York Times*) and a trade publication that specifically focuses on education (*Education Week*), thereby increasing the number of relevant articles included in the dataset. Chapter Three includes additional details about the selection of newspapers for this study.

Studies of Media Coverage of Testing in Schools

Given that education as a broad topic does not receive substantial coverage in the media and that education policy, specifically, receives even less coverage, it is perhaps not surprising that there is scant literature on media coverage of testing in schools and no empirical studies of the framing of the issue by the media. The literature that does exist largely mirrors the studies of education and the media outlined above. That is, this literature has focused primarily on the negative portrayal of educational testing in the media. For example, in an article on improving communication between measurement professionals and the media, Camara and Shaw (2012) argue that the media tend to focus on the problems within educational testing and that this negative focus is compounded by misinterpretations and misunderstandings about concepts and issues that are central to testing.

In an article investigating the role of assessment experts in policy discourse and media coverage of high-stakes testing, Lindle (2009) analyzed *Education Week* and *Washington Post* coverage to gain an understanding of the extent to which assessment expertise was included in political discourse after the passage of NCLB in 2002. Descriptively, Lindle found that political elites and national advocacy groups dominated the media discourse on NCLB assessment policy. In terms of the distribution of quotes and attributes from different sources, political elites accounted for 57% of the sources and advocacy groups accounted for 25%. In contrast, assessment experts accounted for only 10% of the sources (Lindle, 2009). The study also highlighted the discourse concepts in media coverage of NCLB assessment policy (prominent concepts included standards, accountability, adequate yearly progress, and failing schools) and found that assessment experts served two primary functions: 1) as a balancing perspective to political rhetoric about the issue and 2) as interpreters of policy or unintended outcomes (Lindle, 2009).

Assessment scholar Richard Phelps, who has written extensively on high-stakes testing, wrote briefly about the framing of the debate in media coverage. Phelps (2003) argued that media coverage of testing has been heavily biased in favor of critics of testing and that journalists "voluntarily choose to censor themselves" (2003, p. 147), frequently only telling one side of the story. Based on Phelps' argument, the study of the framing of testing in media coverage might be expected to reveal a news slant in favor of opponents of testing.

In a book length treatment on the proliferation of criticism of high-stakes testing and the arguments and rhetoric of this criticism, Phelps (2003) used the metaphor of war to describe the attacks on testing in recent years. Phelps described the strategies and arguments used by testing opponents, grouping arguments into four general areas: standardized tests are 1) not natural, 2)

do not work, 3) are not fair, and 4) are particularly problematic when they are multiple choice. Attack strategies of critics of testing were similarly categorized and include concealing self-interest, ignoring available alternatives, and demonizing the opposition (Phelps, 2003). Included in the book is an "anti-testing vocabulary" which outlines and explains terms used by testing opponents.

In the current study, I am interested in looking deeper into the framing of the issue by tracing the evolution of the dimensions of the issue that have salience over time, regardless of a possible bias in coverage, thereby highlighting how the framing of the issue has changed over time to include different combinations of dimensions. Regardless, this potential for a news slant in coverage is important to keep in mind in the context of the current study.

In addition to these works, there are several studies of media coverage of testing from outside the U.S. In a study of testing coverage in the UK, Warmington and Murphy (2004) identified several "news templates" that shaped coverage of the issue. News templates are defined as "the structural, narrative, and technical formats that exist *prior* to the emergence of specific news events and which are drawn upon by news media in order to produce news issues" (Warmington & Murphy, 2004, p. 289). The authors found that the pervasive narrative in media coverage was the "falling standards" debate and that this became a dominant news template for subsequent coverage of the issue.

Stack (2006) analyzed media coverage by the two national Canadian papers (the *Globe and Mail* and the *National Post*) on the results of the Program for International Student Assessment. The analysis revealed that Canada's relatively high test scores on the international test were used as a proxy for the success of the entire country's education system. The analysis also revealed that a relatively small gap in test score achievement between higher and lower-

income children on the international test (compared to other countries) was used to develop a frame focused on the notion that poverty does not matter for Canada's children as much as it does in other countries (Stack, 2006). Given the relative lack of research on the media's role in shaping policy and public opinion on education, this is an area of research that will benefit from additional investigations.

Significance

As reviewed above, there is a fairly substantial body of literature examining the politics of the testing movement. Similarly, there are numerous studies of the framing of issues in media coverage. However, there are no empirical studies of the specific discourse or arguments used by media in the testing debate, although a few scholars have touched briefly on aspects of this issue. In a series of articles on the politics of testing, McDonnell discussed several of the major purposes of testing used by policymakers and arguments in favor of and against testing (1994, 1997, 2005, 2008). In her work on testing as a policy idea, McDonnell briefly noted that various frames have been used over time to promote test-based accountability, although it is unclear how she identified these frames. A frame from the early 1990s, for example, emphasized the need to improve the nation's economic competitiveness, while a more recent frame from the era of NCLB focused on promoting greater equity in schools (McDonnell, 2004, 2008). As described above, Phelps (2003) also discussed media coverage of the issue but focused exclusively on the apparent bias in news coverage.

To date, no researchers have utilized an empirical approach to identifying the arguments and frames in media coverage of testing or attempted to track the evolution of the debate and the changing nature of media discourse of the issue. The limited literature examining the arguments put forth by proponents and critics of testing in schools is primarily descriptive and the studies

do not employ empirical analytic methods to formally examine the characteristics of these arguments. Additionally, the studies described above do not utilize theoretical frameworks that help to clarify dimensions of arguments. This study adds to the limited literature by conducting an empirical analysis of media coverage of testing over a 20-year period. I employ a conceptual framework based on framing and issue definitions and utilize a recently developed methodology in text analysis called structural topic modeling, which is described in detail in Chapter Three, to identify the various dimensions of the issue and to gain an understanding of whether and how the characterization of this issue has shifted over time.

The substantive purposes of this study are 1) to better understand the framing of the testing debate in media coverage and 2) to identify shifts in the framing of the issue over time. The theoretical and methodological contributions of the study are 1) to examine the utility of structural topic modeling as a methodology for studying issue-definition and framing and for tracking shifts in debates over time, 2) to examine the specific utility of this type of modeling for understanding issues in education policy, and 3) to add to the literature on issue definitions and the impact on public opinion and policy.

CHAPTER 3: METHODS

This chapter describes the methodological approach used in the dissertation to study media coverage of testing in schools. In the following sections, I describe the data sources, the document sample, and the analytic technique. In the first section, I describe the process of selecting appropriate newspapers for the collection of news articles. I then discuss the document sample, including a description of the search terms used to collect news articles and develop the dataset. The third section describes the analytic technique used in this study, beginning with an overview of big data and text analytics, including definitions of key concepts and major assumptions. This is followed by a review of topic modeling methods and a description of the specific technique within this class of methods that I utilize in the study – structural topic modeling.

Data Sources

Selection of Newspapers

The data for my analysis were collected from a trade publication (*Education Week*) and a national newspaper with a general readership (*New York Times*). These news outlets differ in several respects, including their readership, the level of coverage of the issue of testing, and their publication schedule. I chose to include both a trade publication and a general readership newspaper for two reasons. First, I am interested in capturing the potential influence of the media on both the general public and national education politics and on actors in the policy subsystem of K-12 education. Second, including both types of newspaper outlets provides

opportunities for additional analyses exploring the extent to which findings differ across publication source.

Education Week

I collected articles on testing from *Education Week*, a prominent trade publication covering K-12 schooling, to explore coverage of the issue for a professional readership. It is a weekly publication that calls itself "the single 'must read' news source for K-12 leaders and policy experts" (Editorial Projects in Education, n.d.). The readership for *Education Week* includes educators, researchers, and policy actors who are either directly or indirectly involved in K-12 education policy. *Education Week* covers education policy issues (and, specifically, testing) at both the state and national level. For instance, coverage includes reporting on the impact of the reauthorization of the federal Elementary and Secondary Education Act on testing requirements as well as reporting on the state of New Jersey's efforts to include student test scores in teacher evaluations. This breadth of coverage results in a greater number of news articles on testing over the 20-year period than are present in the *New York Times*. See Chapter Four for additional discussion of the dataset and differences by publication.

The *New York Times*

In addition to collecting news articles from *Education Week*, I also collected articles on testing from the *New York Times* to explore national news coverage of the issue. The readership for the *New York Times* likely includes both education stakeholders who are well-informed about education policy issues and a more general audience that may have a less in-depth understanding of these issues. There are several reasons why I selected the *New York Times* as a source for national coverage of the issue instead of alternate newspapers. First, for the purpose of this study, which is to examine trends in coverage over time and how the issue has been framed, the

absolute level of coverage of the issue is not central. It therefore is not essential that level of coverage in the *New York Times* be highly correlated with the level of coverage in other national newspapers. Furthermore, the dimensions of the issue that are prominent at any given time should be similar across media outlets, as should changes in which dimensions are highlighted. Second, I am interested in exploring how frames used by the media differ by whether they are aimed at a general or an education-savvy readership. Finally, the journalistic prominence and breadth and depth of the *New York Times*' coverage are likely to influence the content of other news outlets. Several studies indicate that this is true.

In an extensive treatment of the appropriateness of using the *New York Times* as a representative source for this type of work, Baumgartner, De Boef, and Boydstun (2008) compared the *New York Times* to nine major newspapers[3] to validate their choice of the *Times* to analyze coverage of the death penalty debate. They found that the amount of coverage in the *New York Times* each year corresponded closely with the average amount of coverage in the other nine newspapers (with a correlation of 0.7). When coverage increased in the *New York Times*, it also increased in the other papers (Baumgartner, De Boef, and Boydstun, 2008). Additionally, shifts in the framing of the issue also corresponded closely across the different newspapers. These findings are particularly relevant to the current study in which I develop a similar analysis of shifts in media coverage.

Similar results were found by Boydstun (2013) in an analysis of patterns of coverage of policy issues in the *New York Times*. In the analysis, Boydstun found that patterns of coverage in the *Times* generalized not only to other national newspapers but also to local newspapers,

[3] The nine newspapers were *Boston Globe, Chicago Sun-Times, Denver Post, Houston Chronicle, Los Angeles Times, Miami Herald, Pittsburgh Post-Gazette, Seattle Times*, and *Washington Post*.

television news, and online sources as well (2013). Further support for the *New York Times* serving as an appropriate source for news coverage of an issue, regardless of type of media, is found in a study highlighting the existence of an intermedia agenda setting process, which occurs when prominent news sources influence the content of other news outlets. Golan (2006) found that international news coverage in the *New York Times* was significantly correlated with subsequent international coverage in three television news programs.

This intermedia agenda setting is not only limited to issue coverage but may also apply to salience of issue dimensions. Denham (2014) found that after the *New York Times* introduced new dimensions to coverage of horse racing, these new dimensions became more salient in other media outlets, providing some evidence that the *Times* can influence which dimensions of a given issue are covered.

In an expansive study investigating the conditions under which issue coverage is highly correlated across media outlets and when coverage is likely to be idiosyncratic, Atkinson, Lovett, and Baumgartner (2014) found that issues with high average coverage (whether sustained or periodic) or that exhibit spikes in attention have high correlations across outlets. Conversely, issues with low average coverage and no spikes in attention are not likely to be substantially correlated across outlets. While it is likely that testing will be a high salience issue in a trade publication on education like *Education Week*, coverage in national newspapers like the *New York Times* is likely to be less salient. However, the analysis does suggest spikes in attention in the *Times*, providing further support for the assumption that the *Times* is representative of national coverage. Also, given the literature suggesting that the *Times* is representative of the national media agenda, I proceed with the analysis keeping this possible limitation in mind.

Finally, at a more abstract level than the correspondence in coverage between the *New York Times* and other newspapers, scholars have noted a substantial level of homogeneity of news coverage generally, even in the age of the internet and the explosion of information availability (Denham, 2014). For example, in a study of political blogs and mainstream media coverage of the 2004 presidential election, Lee (2007) found that the media agenda is stable and homogenous across different types of news outlets and regardless of liberal or conservative political leaning. Rogers, Dearing, and Chang (1991) found high correlations between three newspapers (*New York Times*, *Washington Post*, and *Los Angeles Times*) and three television network newscasts (ABC, CBS, and NBC).

Despite the finding of homogeneity in the media agenda across outlets with different political leanings, it remains possible that coverage in the *New York Times* is in some important ways systematically different from other newspapers. Future research might explore the extent to which reporting on testing in schools is homogeneous across different newspapers. For the current study, I am primarily interested in exploring shifts in frames over time and whether differences exist in framing in a general versus a professional news outlet. For these reasons, I limit the scope of my analysis of general readership coverage of the issue of testing to articles from the *New York Times*.

Document Sample

Textual analytic techniques such as topic modeling require a large corpus of documents.[4] In order to ensure that I had a sufficiently large corpus with which to conduct topic modeling and

[4] "Corpus" is a term used to describe a collection of writings, speeches, or other texts. Analysts use corpora to study and describe language.

to also ensure that I captured articles on all aspects of the testing debate, it was necessary to carefully consider the search terms and selection criteria in the collection of news articles.

I developed and used different sets of search terms for each newspaper. This was necessary because *Education Week* is, by default, only reporting on issues related to K-12 schooling. It was not necessary, therefore, to include search terms related to education or schooling. In contrast, a search of the *New York Times* that does not include keywords related to schooling produces results that include testing as it relates to other topics, such as drug testing or DNA testing. ProQuest was used to search the archives of *Education Week* for the keywords "test*," "standardized testing," "high-stakes testing," "high-stakes assessment," "assess*," or "accountability." ProQuest was also used to search the *Times* for the search strings ("school*" OR "education" OR "classroom") AND ("test*" OR "high-stakes testing" OR "assess*" OR "standardized test" OR "accountability"). I limited the search to articles on testing from 1996 through the end of 2015. This 20-year period was chosen to provide a sufficient length of time to track changes in media coverage and to specifically capture multiple years prior to the major high-stakes testing policy change that occurred with the passage of the No Child Left Behind Act in 2001, thereby also allowing for the identification of longer-term trends in the media framing of testing. The searches of ProQuest produced initial results lists of 82,900 articles for *Education Week* and 59,290 articles for the *New York Times* during the 20-year period covered by the study. These initial lists of articles were then further screened for inclusion in the dataset during the data input process. As articles were selected to be downloaded and entered into an Excel database for inclusion in the dataset, I checked article titles for relevance to the study, thereby eliminating a large portion of the articles that were included in the initial, unscreened results lists.

The Analytic Approach: Text Analytics

The structural topic modeling approach I use in this study belongs to a family of techniques developed to analyze large amounts of textual data. In the following sections, I briefly describe the rise of big data, highlight the use of the broad category of methods known as text analytics to address big data, describe topic modeling generally, and explain the specific steps in the structural topic modeling process.

Big Data

In order to understand the growing importance of innovative approaches to analyzing large volumes of data, it is necessary to first explain the rise of big data. The "digital universe" is massive. Enormous amounts of information are produced, collected, and stored each day. The metaphor of a universe is apt: estimates suggest that by 2020, there will be approximately as many bits of information as there are stars in the physical universe (EMC, 2014). The explosion of data is not a recent phenomenon but with advances in personal technology and storage capabilities, the velocity of growth and the diversity of data has changed ("Big Data," n.d.; Friedman, 2012). By some estimates, the amount of data is doubling in size every two years. Accompanying this massive increase in data is a simultaneous increase in computational power. Personal computers now have the capacity to run complex statistical models on huge datasets. With the rise of big data and increases in computational power, there is a concurrent focus in disciplines such as statistics and machine learning on developing new ways of accessing and interacting with these data. Computer scientists, statisticians, and scholars from other fields are beginning to address issues such as how to process and interpret these vast quantities of data. These large volumes of data require innovative and novel statistical techniques, which have developed into the field of big data analytics (King, 2014).

The explosive expansion of the digital universe is greatest for unstructured data (i.e. text, video, and audio). According to a recent report, 90 percent of all data created in the next decade will be unstructured (Gantz & Reinsel, 2011) and these data require different technologies and approaches for analysis. As a result, content analysis techniques for systematically describing texts have evolved to incorporate new technologies and new methods.

Among the various techniques available for conducting content analyses, human coding remains the most widely used. An advantage of human coding is that when multiple coders are used, reliability testing is relatively straightforward. Human coding is also by nature considerably flexible, in that there are few limits on how humans can code data because they not only can read and understand texts but also interpret, for example, beliefs or rhetorical strategies within texts. However, human coding is not without its challenges and is not always practical or feasible. With rapidly growing amounts of publicly available data, it is becoming increasingly necessary to rely on alternative approaches that are less costly and faster than human coding (Nowlin, 2016). As scholars have noted, human coding has great benefits but also several limitations:

> The great benefit of human-coder techniques is that the mapping of words in a text to a topic category is allowed to be highly complicated and contingent. The downside of human coder techniques is that reliability can be a challenge, per-document costs are generally high, and it assumes that both the substance of topics and rules that govern tagging documents with a specific topic are known a priori (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010, p. 210).

Computer-assisted approaches are a viable, powerful alternative to human coding and have been applied productively to unstructured data. Automated text analytics (or text mining) is

the term used to describe the various methods and techniques used to quantify textual data. These analytic methods make possible the systematic analysis of large data collections without the need for significant resources and funding (Grimmer & Stewart, 2013). The dataset for the current study is a large volume of textual data drawn from the 20-year period of media coverage. As an illustration of the size of the dataset, the last three years of coverage (2013 to 2015) in *Education Week* includes over 400 articles each year related to testing. Coding even a single year of these data by hand would be a labor-intensive and time-consuming task. As such, the dataset is well suited to automated text analytics. I use a computer-assisted text analytic technique (structural topic modeling) to answer the research questions outlined in Chapter One.

Compared to other statistical methods, big data techniques such as text analytics are still in their infancy. Because this is a budding field, many of the terms used by researchers working in this area remain ill-defined (Miner et al., 2012). To provide some clarity regarding the key concepts and terms used in this study, I first provide some definitions and make connections between these concepts.

Definitions

Big data analytics. The ability to examine large amounts of data to uncover patterns, correlations, and other insights is called big data analytics. A key feature of big data analytics is the ability to process and analyze data faster and more efficiently than traditional analytic approaches. As a result, big data analytics is increasingly employed in many disciplines. As more and more textual data are digitized, for example, digital humanities research has exploded. For example, a technique called sentiment analysis can be employed to extract and visually represent the emotional trajectory of a novel or movie reviews. In another application, textual analysis has been used for authorship attribution of the Federalist Papers (Jockers & Witten,

2010). Big data analytics is also gaining in popularity within the social sciences, which have also seen an increase in digitized data (Gandomi & Haider, 2015; King, 2014). In subsequent sections, I briefly review the use of text analytics in various disciplines and discuss its application to political science and education, specifically.

Text analytics. At a very general level, the suite of techniques that extract information from textual data to categorize and draw inferences is often referred to as text analytics, text mining, automated content analysis of text, or quantitative text analysis. These techniques were developed, in part, as a way to deal with large text datasets that would be difficult, if not impossible, to analyze using human-based coding (Hopkins & King, 2010). Cluster analysis is the process commonly used in text analytic techniques. Cluster analysis is a broad term to describe the process of identifying similar words or documents to group them together into clusters (Miner et al., 2012; Nowlin, 2016). There are two general approaches to text analytic techniques: supervised learning methods and unsupervised learning methods.

Supervised learning methods. Supervised learning methods are semiautomated. That is, they combine automated processes with some level of human coding or classification. In these models, human coders categorize a subset of documents which are then used to "train" the statistical model how to sort the remaining documents into categories. The key concept in supervised models is that the algorithm in the statistical model learns the proper sorting technique from the human-coded documents (Grimmer & Stewart, 2013). Coding frameworks for these methods, therefore, are developed prior to the computer coding and the models learn only from the human-coded documents.

Unsupervised learning methods. Unlike supervised learning methods, unsupervised methods are not trained using a subset of hand-coded documents. Instead, computers use

properties of the texts to estimate the categorization structure and assign either whole documents or parts of documents to the various categories (Grimmer & Stewart, 2013; Nowlin, 2016). It is then incumbent on the analyst to interpret the output and determine appropriate categories (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). Because a coding framework is not imposed on the data in unsupervised methods, the models provide an opportunity for discovery. That is, unexpected patterns can emerge from the data. One of the popular techniques in unsupervised learning methods is topic modeling. Topic models are used to cluster related words into latent topics without imposing an initial coding or classification structure on the data (Miner et al., 2012). The technique used in the current study is a type of topic modeling called structural topic modeling. I chose an unsupervised learning method for the research in order to explore patterns in the data that may not be accounted for in a supervised approach. That is, there may be frames or frame elements in the dataset that would not be captured using a supervised learning method. I discuss topic modeling and the particular variant used in this study in greater detail later in this chapter.

Some scholars have argued that the major advancements in the application of text analytics in the social sciences have come from unsupervised methods, but that these methods are often more conceptually challenging than supervised modeling (Grimmer & Stewart, 2013) In terms of challenges, it can be particularly difficult to assess the quality of these models because the criteria used to make these assessments are less definitive (DiMaggio, 2015). A discussion of the methods used to assess model validity for topic modeling is included later in this chapter.

Characteristics and Assumptions

There are several characteristics and assumptions that need to be made explicit when using automated text analysis methods. These are drawn from Grimmer and Stewart's (2013) review of text analytics and apply to all approaches to automated text analysis. First, all quantitative models are inherently somewhat inaccurate in the sense that they are unable to account for the full complexity of language. As Grimmer and Stewart (2013) point out, however,

> The data generation process for any text is a mystery - even to linguists. If any one sentence has complicated dependency structure, its meaning could change drastically with the inclusion of new words, and the sentence context could drastically change its meaning (p. 4).

These models, however, are useful for various approaches to classifying or clustering words within documents or documents within corpora.

A second characteristic is that these automated methods are tools that help humans to perform social scientific tasks and cannot serve as a replacement for humans. In both supervised and unsupervised learning methods, human knowledge and interpretation remain at the core of the research enterprise. For supervised methods, the development of the a priori coding framework is critical to a valid analysis. For unsupervised methods, interpretation of output and decisions about key model parameters are critical. The use of one such method – topic modeling – for example, requires a thorough understanding of the topic and the major arguments in the debate in order to interpret and label the latent topics that the model produces.

Third, as with traditional research methods, the choice of model depends on the research questions and the nature of the data. There are no universally applicable models or approaches to

analysis. This is particularly true given that many of these methods are still in their infancy and appropriate (or inappropriate) applications are still being explored and tested.

The final characteristic is that validation is a critical and necessary step in the use of any of these methods. Because these methods are designed to deal with very large corpora and therefore rely on automatic processes, validation provides a check on the output of the models. A discussion of the validation process for the analysis is provided in a later section in this chapter.

Text Analytics and Political Science

The expanding digital universe and the novel methods of big data analytics are beginning to generate significant advances in various subfields of the social sciences, including political science and public policy (Clark & Golder, 2015; King, 2011). The potential of this data-rich future for social sciences is still unclear, given that new approaches are being developed and refined. As Hopkins and King (2010) noted, "Automated content analysis is a new field and is newer still within political science" (p. 230). Reporting on an early study of automated text analysis in political science, Laver, Benoit, and Garry (2003) noted that, at that time, the vast volume of accessible digitized political and policy-relevant text can be liberating for researchers but cautioned that "the big obstacle to this process of liberation is that current techniques of systematic text analysis are very resource intensive, typically involving large amounts of highly skilled labor" (p. 311).

In the past decade, however, scholars have worked across disciplinary boundaries to apply novel text analytic methods that are not as resource intensive to answer important questions about politics and the public policy process. Despite being relatively new, text analytics has been a productive approach to research for political scientists and has the potential

to both advance learning about longstanding theories and produce a broader array of research questions and inferences that were not possible previously (Monroe, Pan, Roberts, Sen, & Sinclair, 2015). There are a growing number of examples of the expansive thinking and analyses that are now possible with big data analytics. For example, scholars are now able to study the purposes of state censorship programs by analyzing censorship on an unprecedented scale. In an analysis of the Chinese government's censorship program, King, Pan, and Roberts (2013) collected over 11 million social media posts from China before the government was able to censor the posts they deemed objectionable. Using text analytic techniques, the researchers were able to compare the content of censored and uncensored posts and found that, contrary to popular belief, the purpose of the censorship program was not to suppress criticism of the state but rather to silence calls for collective action.

One of the big data analytic approaches that has been applied productively to political science is topic modeling. In the next section, I describe topic modeling, including key features and assumptions of the technique.

Topic Modeling

Topic models are a class of unsupervised learning techniques for determining the underlying latent topics or themes in a corpus. These models provide structure to unstructured text collections by identifying patterns of word use and connecting documents with similar patterns (Blei & Lafferty, 2009; Blei, Ng, & Jordan, 2003). David Blei, one of the original developers of the method, describes topic modeling as "statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time" (2012, p. 77).

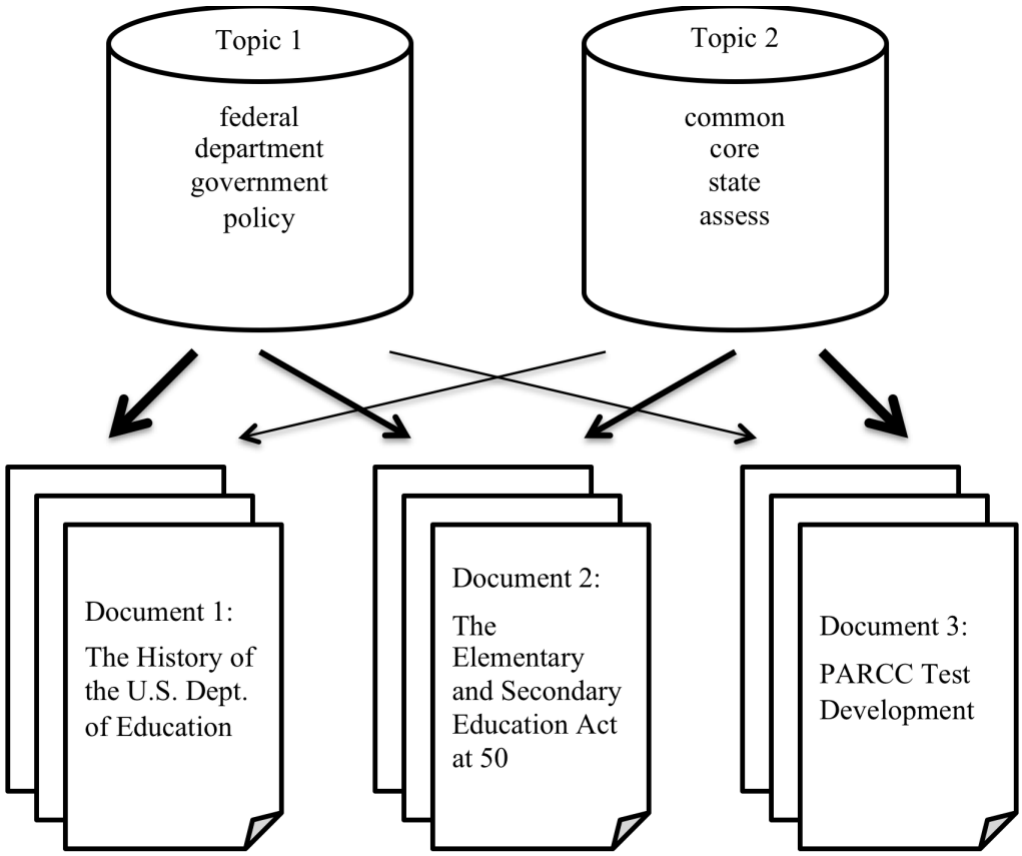
There are several variants of topic modeling. They include, for example, latent Dirichlet allocation (the original topic model), relational topic models (incorporating network analysis), and structural topic modeling. Latent Dirichlet allocation (LDA) was developed by Blei and colleagues in the early 2000s (Blei, Ng, & Jordan, 2003) and remains the most widely used approach. LDA is the form of topic modeling that the variant used in this analysis, structural topic modeling, builds from. In all topic models, topics are distinct concepts that are represented through the words associated with them. For example, in a collection of congressional speeches on transportation policy, one topic may be related to finance, using words such as *investment*, *underfunded*, and *tax*. Another topic may discuss infrastructure disrepair, using words such as *deterioration*, *stress*, and *management*. In the language of topic modeling, the co-occurrence of words across documents in a corpus indicates a high probability of those words being associated with a topic (Grimmer & Stewart, 2013; Roberts, Stewart, & Tingley, 2014). That is, when words cluster together across a set of documents with a frequency that is greater than chance, these co-occurring words constitute an underlying topic. This is essentially how a topic model works to identify topics.

A central feature of topic modeling is the assumption that each document is made up of multiple topics. Because each document contains multiple topics rather than being classified in only one topic area, these models are referred to as *mixed membership models* (Blei, 2012; Grimmer & Stewart, 2013). Additionally, all documents in a corpus are assumed to share the same underlying topic structure. That is, every document in the corpus has an estimated topic proportion for each of the topics in the model. In many cases, this proportion will be either zero or so small that the topic will essentially not be present in the document (Blei, 2012). For example, in a topic model estimating a total of 12 topics, each document in the dataset is

assumed to contain a proportion of each of the 12 topics. However, few, if any, documents are likely to contain all 12 topics. Therefore, most documents will have some topic proportions that are zero.

Another key feature of topic modeling is the ability to not only identify the latent topics in a collection of documents, but also to identify the prevalence of each topic within each individual document. As a simplified example, a collection of documents might have four latent topics: natural disasters, meteorology, disease control, and data analysis. An article about climate change might be a blend of the following topics: natural disasters (with co-occurring words like "flood" and "hurricane"), meteorology (with co-occurring words like "weather" and "atmosphere"), and data analysis (with co-occurring words like "model" and "parameters"). Another article in the same collection might also be about climate change but focus on natural disasters and disease control and these two topics would have greater proportions than the meteorology and data analysis topics. An individual document, then, is a particular collection of words that have an underlying topic distribution (Blei, 2012). Topic modeling looks for all topics in every document, thereby increasing the likelihood of finding topics that may not be the primary subject of a given document but that are present within the document nonetheless. This feature of topic modeling is significant for the current study because it increases the level of confidence that all instances of a given topic within the dataset are accounted for.

Figure 3.1. Illustration of Topic Modeling



Adapted from Underwood, 2012

Figure 3.1 is a simplified graphical representation of how topic modeling works. The two latent topics are represented at the top of the figure and include a few words that make up the topics by co-occurring repeatedly across documents. These topics pertain to education – topic 1 might be labeled *Federal Role in Education* and topic 2 might be labeled *Common Core State Standards*. The thickness of the arrows indicates the probability of a topic being associated with a given document. Thicker lines equal a higher probability (more of the associated words appear together in a given document). In this example, topic 1 has a high probability of being associated with document 1 about the history of the U.S. Department of Education but a lower

probability of association with document 3 about PARCC tests. This suggests that document 1 is primarily about the federal role in education. Topic 2, on the other hand, has a high probability of being associated with document 3 and a low probability of association with document 1. Document 3, therefore, is primarily concerned with the Common Core State Standards. The two topics have a similar probability of being associated with document 2, which indicates that this document is about both topics.

Because topic modeling is an unsupervised learning method, it is useful as a tool for discovering new patterns or topics through exploration of a collection of documents. For example, as a scholar of education policy, I would expect a topic model of documents about NCLB to estimate certain topics such as accountability, testing, achievement gap, and so on. But without imposing this topical structure on the corpus a priori, the model allows for other topics to emerge from the data, topics that may be different from or in addition to those an analyst would hypothesize would be included.

There are numerous reviews of this technique available online and in print, from Matthew Jockers' (2014) user-friendly narrative introduction to the concepts to Megan Brett's (2012) overview in the Journal of Digital Humanities to David Blei's (2012) more technical explanation. Variations on topic modeling that have been used in political science include dynamic topic modeling (Quinn et al, 2010), expressed agenda modeling (Grimmer, 2010), and structural topic modeling (Roberts, Stewart, & Airoldi, 2015). These variations use the same basic structure as LDA.

When applied to research on issue definitions and framing, topic modeling provides a method for identifying the different frame elements of an issue that contribute to different frames. Topic modeling can be used to analyze collections of documents at different levels of

granularity. A topic model on presidential state of the union addresses, for example, would likely generate a broad array of topics on issues as varied as health care, education, and foreign policy. In contrast, the topic model that I ran on media coverage of testing in education generated a substantially more targeted set of topics related to testing and schooling. An example of the application of topic modeling to identify issue dimensions within a single policy debate is Nowlin's (2016) study of used nuclear fuel policy. In this study, the author used topic modeling to estimate seven distinct dimensions in 140 Congressional hearings over a 38-year period in the debate over management of used nuclear fuel. The seven dimensions also varied in terms of the proportion of discussion paid to each over the period of the study. This specific type of application most closely aligns with the current study, in which I employ topic modeling to identify issue dimensions within the testing policy debate.

Some illustrative applications of topic modeling from different disciplines will illuminate the breadth of application of the technique and help to further explicate the utility of the method. In health care, for example, several studies have employed topic modeling to develop mortality prediction models in intensive care units (Ghassemi, Naumann, Joshi, & Rumshisky, 2012; Jo, Loghmanpour, & Rose, 2015). By providing an efficient way to analyze large volumes of written provider notes and reports, the researchers were able to identify which topics were associated with different health outcomes, including in-hospital mortality and long-term survivors. This new method for analyzing data to identify conditions associated with mortality outcomes could be used to improve ICU care.

Topic modeling has also been productively applied to answer research questions in the humanities. Jockers and Mimno (2013), for example, used topic modeling to identify themes in novels. This approach has benefits over traditional methods for assessing thematic content in

novels in that it eliminates individual reader bias. The method also expands the analysis of thematic content to large-scale projects, which was previously limited by what a scholar had time to read and interpret. In this case, topic modeling allowed for the inclusion of over 3,200 works of 19th century fiction. Topic modeling not only identified the themes across the corpus but also revealed insights such as the extent to which male and female authors use different themes in their writing.

Historical analyses have also been developed using topic modeling. For example, the technique was used to examine nearly the entire archive of the *Richmond Daily Dispatch* (over 112,000 pieces), which was published from 1860 to 1865, providing a rich account of Civil War-era Richmond (Nelson, 2011). The study revealed, for example, articles associated with topics such as "Anti-Northern diatribes," which argued that it was righteous to kill Northerners because they were not true Christians, as well as other insights about the arguments and appeals that were used to encourage men to fight in the war (Nelson, 2011).

In an example from political science, topic modeling was applied to an analysis of more than 24,000 Senate press releases from 2007 to identify how politicians articulate their priorities to constituents (Grimmer, 2010). In the study, 43 topics were estimated, such as "food safety," "illegal immigration," and "energy policy." The author concluded that topic modeling is useful for answering theoretically important questions about political communication.

While many disciplines have begun to productively explore the potential of automated content analysis of text, education researchers have made few ventures into this area (Bowers & Chen, 2015). In a review of approaches to automated content analysis of text, for example, O'Connor, Bamman, and Smith (2011) highlight fields in which this method has been applied, including political science, economics, psychology, sociolinguistics, public health, history, and

literature. Absent from this list is education. This is beginning to change, however, as a few education studies have been published recently that utilize this technique.

In a longitudinal investigation of school district capital facility bond election proposals, Bowers and Chen (2015) applied a topic model to analyze the full text of over 1,200 proposals to identify bond proposal topics and then examined the probability of passage and voter turnout based on topics. The authors identified nine topics across the bond proposals. These nine topics were then used as independent variables to examine the relationship between bond proposal topic and passage. Findings suggested that bonds focused on athletic facilities had a lower probability of passing than all other bond topics.

Topic modeling has also been applied to massive open online courses (MOOCs) to identify patterns in students' texts, including course evaluations and online forum discussions. This application is particularly relevant given that MOOCs often enroll tens of thousands of students, making the process of reading and analyzing student feedback or discussion forums overwhelming. Using structural topic modeling, Reich, Tingley, Leder-Luis, Roberts, and Stewart (2015) analyzed data from MOOCs to determine patterns in students' written responses in discussion forums and course evaluations. The use of structural topic modeling also allowed the researchers to describe how student responses varied by covariates such as level of satisfaction with the online course. In another application, topic modeling was used to identify patterns and evaluate pre-service teachers' reflective journal writings (Chen, Yu, Zhang, & Yu, 2016).

Similar to other disciplines, large collections of digitized textual data are becoming more widespread in education. As education researchers begin to wrestle with analyzing these large

datasets, topic modeling and other forms of text analytics will be useful tools for answering emerging research questions.

In each of the studies described above, scholars leveraged topic modeling to analyze large volumes of text that would otherwise require a level of coding and interpreting that would be nearly impossible if not computer-assisted and were able to pose (and answer) important and fascinating research questions that might not otherwise be feasible to explore. As productive as topic modeling has been in recent years across many disciplines, education stands out for a relative lack of applications. The current study seeks to add to this limited literature and provides a test of the extent to which topic modeling can be productively applied to media coverage of education policies. In addition to exploring the use of topic modeling in this domain, this technique is particularly appropriate for answering my research questions. I am interested in identifying and tracking the evolution of frames in the testing debate over a 20-year period and across a large volume of news articles. Topic modeling will assist me in identifying frame elements and in measuring salience, resonance, and persistence of frames. In the next section, I describe structural topic models, the topic modeling technique used in the current study. I then detail the steps in the analytic process.

Structural Topic Modeling

My analytic approach draws on the work done in the methodological area described above (text analytics) as well as the conceptual framework of framing and the relationship between media and the public and policymaking. Specifically, I employ a recently developed unsupervised learning method called *structural topic modeling* (STM) to identify the frames in the high-stakes testing debate. Structural topic modeling is a specific type of topic modeling that includes document-level metadata (such as author, publication date, or source) to facilitate

hypothesis testing between the topics and the document-level information, thus allowing for modeling of topics that accounts for the nature of the documents (Roberts, Stewart, & Airoldi, 2015; Roberts, Stewart, & Tingley, 2014).

The primary innovation of structural topic modeling, and what distinguishes it from other topic modeling techniques, is that it permits analysts to include information about each document (document-level metadata or covariates) into the topic model. Including document-level metadata allows for insights into the topical prevalence or topical content of documents (Lucas et al., 2014). Topical prevalence is how much of a document is associated with a topic and topical content is the words used within a topic. So, for example, a researcher might be interested in the extent to which conservative news sources and liberal news sources use similar or different language in their coverage of the Black Lives Matter movement. By including a political ideology covariate in the STM, this type of analysis is possible by examining differences in topical content across the different news sources.

For my purposes, the inclusion of document-level metadata allows me to pool the data from the two news sources and run analyses that look across all articles in the corpus but also allows me to explore the data by the different news sources to gain an understanding of the extent to which the framing of the issue is different across news outlets. The statistical programming software *R* and the "stm" package developed by Roberts, Stewart, and Tingley (2014) are used in the analysis.

Regardless of what is being studied or the analytic approach employed, transparency in the steps taken and decisions made is critical. This is particularly important for novel techniques such as topic modeling and other text analytic methods, as these methods create unique

challenges to transparency and replication (Romney, Stewart, & Tingley, 2015). In an effort to address transparency, I outline the specific steps in the analytic process in the following sections.

Steps in the Analysis

Pre-processing steps. Text mining techniques require a series of pre-processing steps that prepare the data for modeling (see Newman & Block, 2006 for an example description). The first step is to remove numbers and punctuation from the corpus and make all terms lowercase. The next step is to remove stop words from the corpus. Stop words are words that do not contribute to the interpretation of the estimated topics and therefore should be filtered out before running models. Standard stop word lists include common words such as "and," "the," and "this." After conducting pilot models with custom lists of additional stop words (words that might be common across many of the articles on testing), I determined that the stop word list included in the "stm" package in *R* was sufficient. The next step is to stem all of the terms in the corpus. Stemming removes the suffix of words in order to equate similar terms (e.g. "test" stands for test or testing; "technolog" stands for technology or technologies). These pre-processing steps are necessary not only for developing a dataset that is formatted properly for conducting topic modeling but also to aid in the interpretability of the output of latent topics.

Selection of number of topics. Determining the appropriate number of topics is a critical step in developing the model (AlSumait, Barbara, Gentle, & Domeniconi, 2009; Grimmer & Stewart, 2013). The appropriate number of topics can be difficult to determine in the initial development of the model given that unsupervised learning methods have as a key assumption that classification into topics is determined by the nature of the corpus itself. It is also somewhat arbitrary in that researchers can manipulate the number of topics relatively easily in the model specification. However, topics must have semantic validity. That is, they must be discernible

based on the words that are highly associated with the topic. This is also a key step in the process of validating the model, described below. Being knowledgeable about the policy area under investigation is helpful for determining the initial parameter for the model on the number of latent topics and for assessing semantic validity. I will return to this issue of selection of the number of topics in a later section of this chapter on process checks.

Modeling and interpreting topics. Modeling is done in *R* using the "stm" package. Structural topic models take advantage of included covariates to allow for more detailed analysis, such as an investigation of how frame salience differs by publication source. I include document-level metadata on source (either *Education Week* or *New York Times*) as well as date of publication in the dataset for the model. The first model was set to estimate 20 topics and the final model used in the analysis was set to estimate 50 topics (a more detailed discussion of developing the model and determining the appropriate number of topics is included in Chapter Four). In addition to estimating topics, the output of structural topic modeling also includes *highest probability* words and *frequent and exclusive* words for each topic. Highest probability words are words that the model estimates to have the highest probability of being associated with each topic. Frequent and exclusive words are words from the dataset that also have a high probability of being associated with a topic and that also are unique to the given topic. For example, an estimated topic from a pilot model (run with a small subset of data) had the following output:

Highest probability words: *standard*, *federal*, *department*, *state*

Frequent and exclusive words: *duncan*, *waiver*, *nclb*, *republican*

The interpretation of topics begins with a visual review of the terms that are associated with each topic. A likely initial interpretation of the example above is that the latent topic is about the

federal role in testing policy. Furthermore, the frequent and exclusive terms suggest that the topic is about a specific aspect of the federal role, namely the No Child Left Behind Act (*waivers* refers to the granting of flexibility by the U.S. DOE on certain provisions of the law). After reviewing the lists of terms associated with each topic, I examined a sample of articles from the corpus that exhibit the topic with a high probability. For example, articles in which words are most likely to be associated with the example topic above should be primarily concerned with the federal role in testing policy. The *stm* package in R also provides options for creating visuals of the output of the topic model. The figures included in Chapter Four were generated from commands in the *stm* package.

Process checks/validation. One of the major concerns about unsupervised learning methods such as topic modeling is validation. With the rise of big data analytics, researchers are beginning to study the extent of agreement between human coding of texts and automated coding techniques, and results suggest that automatic coding is often closely aligned with human coding (Albaugh et al., 2014; Tanata, Hallgren, Imel, Atkins, & Srikumar, 2016). However, model validation remains a critical step in topic modeling.

Scholars working on developing and refining unsupervised learning methods such as topic modeling have emphasized the importance of validating models (Chang, Boyd-Graber, Wang, Gerrish, & Blei, 2009; Grimmer & Stewart, 2013; McFarland et al., 2013; Nowlin, 2016). To a large extent, the validation process focuses on determining the correct number of latent topics in the model. Too few or too many topics can substantially alter the interpretation of the results. Several methods for measuring topic quality are explicated in the literature. One approach is to assess the relevance of each topic to the corpus (McFarland et al., 2013). For example, a corpus of sports medicine-related articles might generate some latent topics related to

other fields of medicine that are not relevant to the analysis, indicating that the search terms picked up articles with keywords on medicine that are not sports-specific. This check on topic relevance also provides a way to screen out irrelevant articles that have been inadvertently included in searches that cast a wide net for documents. Another technique is to measure topic coherence or entropy. Topics with high entropy are likely to be noise rather than meaningful estimates of a latent category. These topics are also more difficult to interpret and may need to be excluded from the analysis.

Another technique for assessing model and topic quality is predictive validity (Grimmer, 2010; Grimmer & Stewart, 2013; Quinn et al., 2010). This technique (which is similar to construct validity) is based on the theoretical work on the relationship between the media and real world events or objective conditions (Behr & Iyengar, 1985). Because the media respond to external events, these events should explain spikes in attention to a given topic. This can be assessed graphically using plots of changes in topic attention over time. Spikes in attention should map closely to external events relevant to the given topic. As an illustration, if I labeled a latent topic in the testing coverage as "federal role in education," I might expect to see a spike in attention to this topic around the time of the passage of NCLB. This would support the validity of the accountability topic. As part of my analysis, I will develop a timeline of external events and policy developments related to the issue of high-stakes testing that will be used to assess the validity of the model and the topics.

Sensitivity analysis can be used to iterate on the process of modeling by varying the number of topics and examining each iteration. Too few topics will result in overlapping terms and distinct topics will be more challenging to identify. With too many topics in the model, there may be clusters that lack semantic validity or that could be combined under a broader topic

(Nowlin, 2016). It is important to emphasize that there is no "right" number of topics to include in the model. Instead, determining the appropriate number is highly dependent on the research questions and the results of sensitivity analyses and process checks.

In an effort to be both transparent and thorough, I conducted a series of validity analyses in the current study, using the techniques described above. During the modeling process, I conducted sensitivity analyses by iterating on the number of topics in the model parameters. Because this study was designed to identify all relevant media coverage of the testing issue, I started the model iterations by specifying a relatively high number of topics for the model. I also started with a high number of topics so that irrelevant groupings of words would be captured in distinct "undefined" topics. Specifying too few topics in the model can lead to uninterpretable topics when irrelevant words obfuscate the topic meaning. After multiple iterations, I settled on a final model by assessing topic relevance to the corpus and examining topics for coherence. Finally, I measured predictive validity by examining the relationship between topic attention over time and external events by mapping topics to a chronology of relevant policies and events in the testing debate. I discuss this mapping in the next section.

Mapping Topics to Policy Developments

In addition to identifying shifts in the media coverage of high-stakes testing, I am also interested in exploring whether and how these shifts relate to changes in policy on the issue. I developed a chronology of noteworthy policies and events related to high-stakes testing and accountability in K-12 public schooling. This chronology serves two purposes. First, it provides historical context for the debate and serves as a review of the major developments in this policy area over the past 20 years. By overlaying the evolving framing of the issue with the chronology of testing, I am able to begin to explore possible connections between media coverage and

framing and policy developments. Second, the chronology provides a more complete picture of the debate than is feasible by only focusing on the media framing of the issue, which is important for understanding the extent to which this issue has evolved as a result of numerous individual elements working independently but toward similar goals to shift the issue definition (see, for example, Baumgartner, De Boef, and Boydstun, 2008).

CHAPTER 4: FINDINGS

This chapter describes the findings of the analysis using topic modeling. In the first section, I provide descriptive statistics of the dataset. I then discuss the processes of determining the appropriate number of topics to include in the topic model and determining appropriate labels for each topic, as well as details of the top topics. The third section is a discussion of the changes in topic coverage over the study period (1996 through 2015). I then detail whether and how topic coverage is similar or different by publication source, comparing coverage in the *New York Times* (a general readership newspaper) with coverage in *Education Week* (a professional readership newspaper). The following section describes the frames that emerge from the combinations of salient frame elements (topics) across time. Finally, I overlay the frames with a timeline of testing policies and events to explore possible relationships between media coverage and policy developments.

Descriptive Statistics

The total number of articles included in the dataset was 8,161 (see Table 4.1). There was a total of 2,280,048 words or tokens in the dataset (words are often referred to as tokens in the text mining literature). Of the 52,128 terms (nonrepeating words or vocabulary entries) in these documents, 27,513 were removed due to high frequency across all the documents, leaving a total of 24,615 terms in the dataset. *Education Week* had a higher total number of articles on testing (5,196) during this 20-year period than the *New York Times* (2,965). The higher number of articles in *Education Week* was expected, given that the publication focuses specifically on education, which the *New York Times* does not. The relatively high number of articles in the

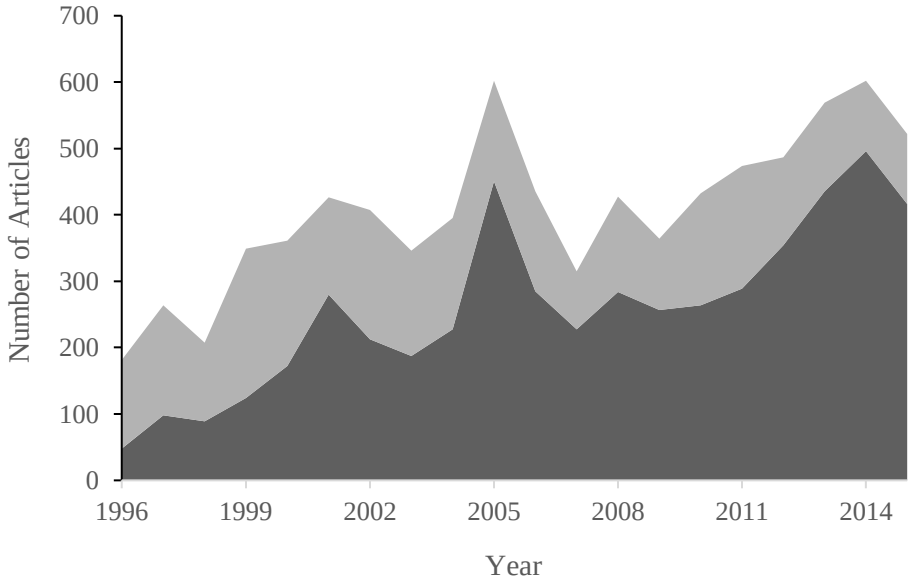
New York Times with coverage of educational testing (compared to coverage in *Education Week*) was surprising but may reflect that the *New York Times* is a daily newspaper and *Education Week* is a weekly newspaper. Almost 3,000 articles in a 20-year period is an average of about 150 articles each year that include at least some coverage of testing in schools (almost three stories per week).

As shown in Figure 4.1, the range for articles across the two publications by year was [181, 602] with an average of 408 articles on testing per year. There also appears to be an overall increase in coverage of testing over time. For example, a total of 1,360 articles related to testing were included in the first five years of the dataset (1996 to 2000) and a total of 2,654 articles were included in the last five years (2011 to 2015). This increase in overall coverage of testing was primarily driven by the increase in coverage in *Education Week*, which included 531 articles in the first five years of the dataset and 1,990 articles in the last five years (a 275% increase). Coverage in the *New York Times* was much more consistent across time. The dramatic increase in coverage of testing in *Education Week* may reflect a growing attention to the issue as testing gained prominence through state and federal policies and became increasingly controversial.

Table 4.1. Description of the Dataset

Documents	
Education Week	5,196
New York Times	2,965
Total Articles	8,161
Words	
Tokens	2,280,048
Terms	24,615

Figure 4.1. Frequencies for Newspaper Articles by Year



Determining the Appropriate Number of Topics

In this section, I explain the process for determining an appropriate number of topics to specify in the topic model parameters (in the *stm* package in R the estimated topic number is labeled *k*). This was an iterative process that required specifying several different models with *k* varying in each model. I began with k=40, then ran models with k=20, k=30, k=45, k=50, and k=60. In general, the models converged after about 150 to 200 iterations in R. In assessing the fit of the models and determining the final model, I compared the outputs for each model with varying *k* to find the model that had both a relatively low number of undefined topics while also maintaining topic coherence for the labeled topics. As noted in Chapter Three, this aspect of the process is subjective and there is no "correct" number of topics for any issue. Rather, the iterative process of running and analyzing multiple models is used to arrive at a final model that the analyst determines is appropriate for further analysis.

The final model included a 50-topic specification (k=50) with 12 of those 50 topics remaining undefined for a total of 38 topics. Topics labeled as "undefined" were topics in the model output that lacked semantic validity (i.e., did not have an interpretable meaning based on the co-occurring terms highly associated with those topics). For example, topic 10 included the following associated terms: *say, get, like, one, just, class, can, homework, film, laptop, wear, fun, movi, desk*. These terms are uninterpretable as a topic related to testing and therefore topic 10 was labeled as undefined. All similarly uninterpretable topics were excluded from the analysis. A final count of 38 topics indicates that testing is a pervasive issue in education that was discussed in a diverse array of contexts. That is, as a core component of schooling, testing was an element of the coverage of many other issues in education, including topics as varied as teacher merit pay, gender disparities, and English Language Learners, for example.

As suggested in the structural topic modeling literature, exploration of the estimated topics should include both an examination of the collection of words associated with each topic and an examination of articles highly associated with each topic (Roberts, Stewart & Tingley, 2014). For the topics estimated in my final model, the collections of words that were highly associated with topics were often sufficient to indicate appropriate labels for each topic even before an examination of specific articles associated with each topic. However, a few topics lacked enough specificity in the highly-associated words and required further investigation of associated articles. Two sets of terms are included in the model output for each topic: highest probability words and frequent and exclusive words (FREX). Highest probability words are words that have the highest probability of being associated with a given topic, based on the co-occurrence of words as described in Chapter Three. Frequent and exclusive words are words that are weighted by their frequency and exclusivity to a given topic. For example, one topic had the

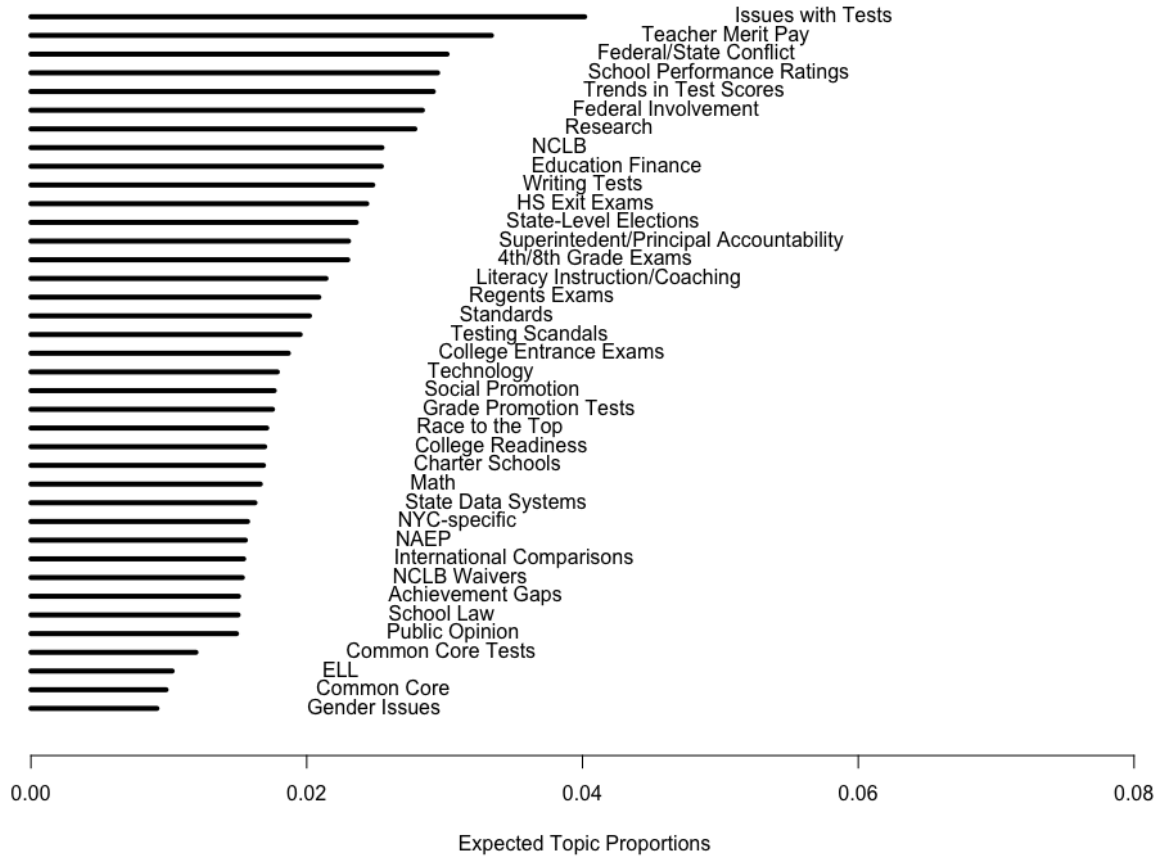
following highest probability word stems: *nation, naep, assess, read, profici, progress, math*. These word stems suggest that the topic is about the National Assessment of Educational Progress (NAEP). I confirmed that NAEP was the correct label for this topic by analyzing several articles that were highly associated with the topic. These articles had titles such as "Panel to Study NAEP Use as Check on Tests" and "NAEP's Board Mulls Five-State Report." All defined topics had correspondence between associated words and articles. That is, both the associated words and the associated articles for each topic provided evidence for the same understanding (and therefore labeling) of the topic. Details for each of the 38 topics are included in the Appendix.

Topics in the Testing Debate

The final model used in the analysis had a total of 38 defined topics. As described in Chapter Three, each topic in the model has an estimated proportion assigned to each individual article. That is, the model estimates a percentage of each article that is dedicated to each topic (in the modeling process, individual articles are assigned a percentage of every topic, but many of these percentages will be zero). Each topic also has an overall proportion across the corpus of articles included in the 20-year period. For each article, the topic proportions sum to 100 percent and the total topic proportions across the entire corpus also sum to 100 percent. Higher overall proportions in the corpus indicate higher salience across the full span of articles included in the dataset. Therefore, the top topics in the dataset represent issues related to testing that were prevalent in media coverage for substantial portions of the 20-year period. Conversely, topics near the bottom of the dataset in terms of proportional representation were less prevalent in media coverage. Figure 4.2 illustrates the proportional representation of each topic across the entire corpus of articles. As shown in the figure, proportional representation ranged from a

maximum of approximately 4 percent (the Issues with Testing topic) to a minimum of approximately 1 percent (the Gender Issues topic). To further highlight the relationship between topics and terms, in the next section I provide greater detail on the ten topics with the largest proportional representation in the corpus.

Figure 4.2. Topics by Proportion



Top Topics

The ten topics with the highest proportional representation in the corpus are listed in Table 4.2 below, along with the ten terms that had the highest probability of being associated with each topic and ten frequent and exclusive terms for each topic. A complete table with all 38 topics is included in the Appendix.

Table 4.2. Top Topics by Proportion and Terms Associated with Each Topic

Rank	Topic	Terms	Frequent, Exclusive Terms
1	Issues with Tests	test, score, assess, student, exam, use, measur, take, result, administ	test, error, fairtest, highstak, administ, valid, ctbmcgrawhil, diagnost, stake, reliabl
2	Teacher Merit Pay	teacher, union, evalu, teach, pay, system, year, perform, effect, bonus	bonus, salari, certif, union, licens, compens, teacher, profess, tenur, incent
3	Federal/State Conflict	state, feder, depart, educ, requir, offici, year, district, must, meet	connecticut, regul, compli, subgroup, titl, utah, depart, adequ, complianc, feder
4	School Performance Ratings	school, improv, account, system, achiev, perform, student, state, progress, measur	account, improv, card, target, school, lowperform, perform, goal, achiev, measur
5	Trends in Test Scores	score, point, year, read, averag, math, gain, percent, result, show	percentag, gain, averag, rose, slight, percentil, proport, point, declin, score
6	Federal Involvement	educ, bill, hous, bush, senat, republican, presid, democrat, committe, propos	sen, senat, mccain, rep, bipartisan, goodl, bill, hous, bush, esea
7	Research	studi, research, found, univers, effect, find, professor, differ, result, educ	valuead, research, studi, random, stanford, found, conclud, conclus, effect, rand
8	No Child Left Behind	law, child, left, behind, feder, educ, act, nclb, secretari, spell	nclb, ayp, law, spell, left, behind, paig, child, margaret, jen
9	Education Finance	money, fund, budget, program, spend, million, year, voucher, increas, cost	voucher, tax, budget, fiscal, tuition, spend, cut, money, fund, financ
10	Writing Tests	student, question, write, skill, use, assess, answer, can, ask, knowledg	format, vocabulari, reader, passag, write, essay, cognit, skill, grammar, literatur

Note: One of the pre-processing steps to prepare the data for modeling is stemming all words to equate similar terms (e.g. “averag” stands for average, averages, and averaging; see Chapter Three).

Arguably, these top topics include some of the more controversial issues associated with testing, including issues with tests such as validity and test errors, school performance ratings, and teacher merit pay. Many of these top topics not only had a high overall proportion in the dataset across the 20-year period but also had relatively high proportions at specific time periods through the 20 years, which is discussed in the next section. These top ten topics are briefly described below.

- *Issues with Testing*: The Issues with Testing topic deals largely with the potential problems of tests, including test accuracy and test misuse. Important terms include "measur," "error," and "valid." Articles highly associated with this topic include "Stanford Report Questions Accuracy of Tests," "School Board in New Dispute on Test Scores," and "Educators' Indifference to The Misuse of Standardized Tests is Having Calamitous Consequences."
- *Teacher Merit Pay*: This topic is about teacher pay and deals with the issue of merit pay for teachers based largely on student test scores, as evidenced by terms such as "teacher," "pay," "perform," and "salari." Associated articles include ""Union Signals Softer Stance on Merit Pay," and "Model Plan of Merit Pay in Ferment".
- *Federal/State Conflict*: The Federal/State Conflict topic is primarily concerned with tensions between federal accountability and testing policy and state responses. This is seen in terms such as "state," "federal," "complianc," "connecticut," and "utah." Connecticut and Utah were two states that gained media attention for explicitly challenging the requirements of NCLB. Articles highly associated with this topic often described tensions between the U.S. Department of Education and states and

include headlines such as “Department Sets Timing for Accountability Plans,” and “Department Raps States on Testing.”

- *School Performance Ratings*: The School Performance Ratings topic deals with policies and processes to give schools grades or other performance ratings that are largely based on student test scores. Important terms include “improv,” “account,” “progress,” and “target” and articles include “For High Schools, A's and Low Grades Rise,” “Letter Grades Look Simple, But Realities are Complex,” and “Calif. Schools Get Rankings Based on Tests.”
- *Trends in Test Scores*: The Trends in Test Scores topic describes coverage of school and district performance on tests and includes terms such as “score,” “percentil,” “gain,” and “declin.” Example articles include “NAEP Reports Modest Gains in Math and Science Scores,” “SAT Scores Hold Steady for '09, Panel Says,” and “Test Scores Still on Upswing in Urban School Districts, Report Finds.”
- *Federal Involvement*: The Federal Involvement topic primarily covers two major aspects of the federal government’s involvement in education: ESEA reauthorization and discussions of a national test. Terms include “hous,” “senat,” “committe,” and “esea” and associated articles include “Deal on National Test Faces Opposition from All Sides,” “ESEA Bill on Track as Senate Changes Hands,” and “Education Bills on Congress' Fall Agenda.”
- *Research*: This topic primarily deals with coverage of research and studies using test score data. Highly associated terms include “studi,” “research,” “effect,” and “result,” and articles include “Schoolwide Reform Improves Test Scores, Analysis Finds” and “Study of Reading Programs Finds Little Proof of Gains in Student

Comprehension."

- *No Child Left Behind:* This topic is about the federal No Child Left Behind Act and issues related to the testing mandates of the law, as supported by terms such as "law," child," "feder," and "nclb" and articles including "Flaw in 'No Child' Law" and "Test Standards Cut as Sanctions Loom."
- *Education Finance:* The Education Finance topic includes coverage of issues related to spending and budgeting in education and also includes issues concerning the financial implications of vouchers, as indicated by terms such as "fund," "budget," "fiscal," and "voucher." Associated articles include "Florida Districts Slash Programs, Personnel," "Vouchers Prove Wild Card for Local Finances," and "Iowa: Teacher-Salary Package Cuts a New Policy Path."
- *Writing Tests:* This topic deals with issues related to writing and writing tests as evidenced by terms such as "question," "write," "vocabulari," and "grammar." Articles associated with this topic include "Teaching Writing Involves Other Subjects as Well" and "Modifying the Subject."

Additionally, individual topics varied over time in their proportion of coverage in the corpus. This is illustrated in Figure 4.3, which shows topic proportions across time for all 38 defined topics. Many of the topics cluster within the range from 1.5 to 2.5 percent in proportional coverage. However, some topics exhibit spikes in coverage at different periods of time. For example, a topic on the Common Core State Standards had essentially zero coverage in the dataset until 2008 (the years prior to the introduction and adoption of the standards) but in the final years of the dataset increased substantially in the amount of coverage (see Figure 4.4). Because this topic was not present in the data for the first 12 years, it is near the bottom in Figure

4.2, which represents overall proportional coverage in the dataset. However, this topic began to receive media coverage in the final years of the study, particularly from 2012 through 2014. As illustrated in Figure 4.3, some topics exhibited substantial variation in coverage across time and a subset of topics peaked in the five to six percent range in proportionality. The use of topic salience to determine frame elements is discussed in the following section.

Figure 4.3. Changes in Topic Proportions Across Time, 1996-2015

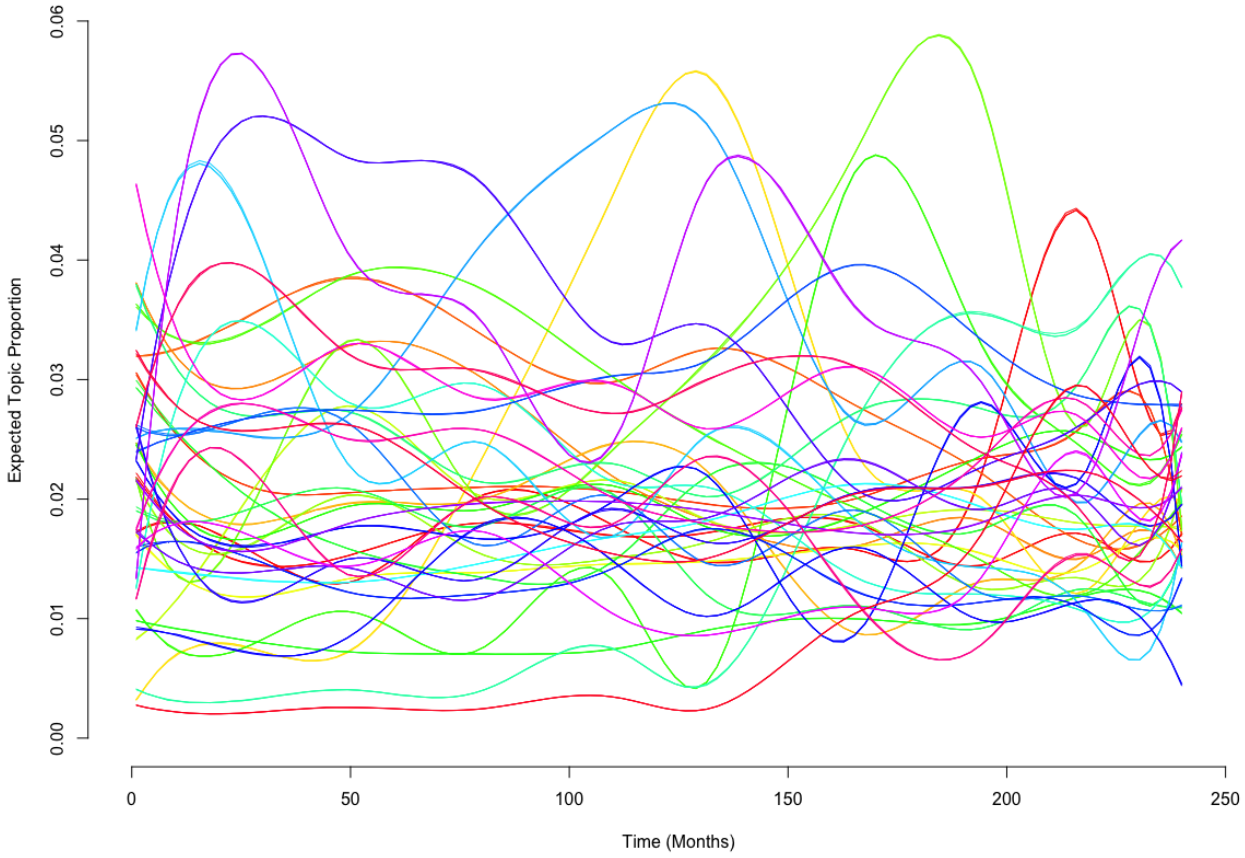
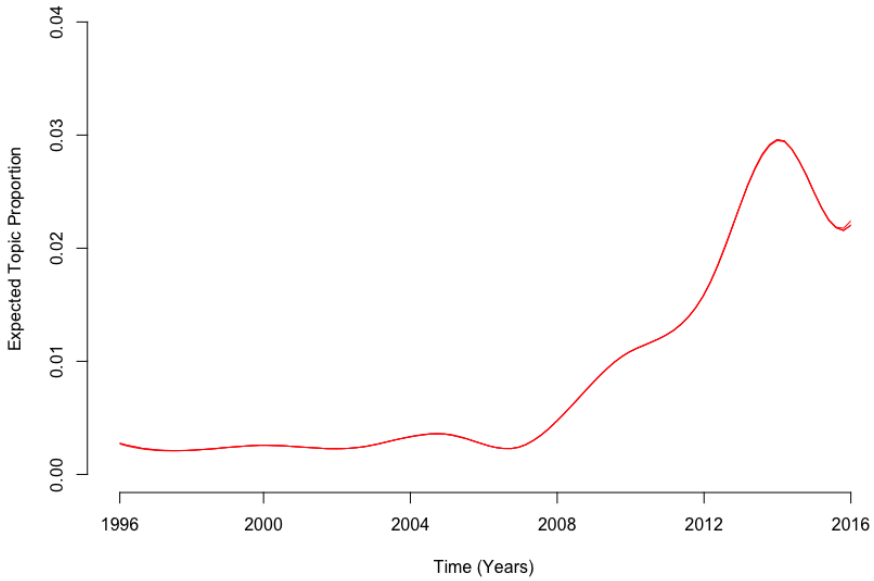


Figure 4.4. Coverage of the Common Core Controversy



Determining Frame Elements from Salient Topics

As described in Chapter Two, I employ Entman's (1993) definition of framing, which is a process of selecting and highlighting certain aspects of an issue. I also apply Matthes and Kohring's (2008) work on measuring frames, in which they argue that frame elements are less difficult to identify than frames. Once frame elements are identified, frames can be identified by analyzing how frame elements cluster together. In the current study, topics are the labeled clusters of co-occurring words that result from the topic modeling process. When these topics were particularly salient, I labeled them as frame elements. Frame elements cluster together to make up frames, which are ways that an issue is understood. Frame elements can also be labeled as positive, negative, or neutral, and this labeling can provide insights into the overall tone of framing of an issue. When frame elements change (elements become more or less salient), the underlying frame changes and, therefore, the understanding of the issue changes. Utilizing this conception of how framing works, I analyzed the results from the topic model estimation to determine frame elements. I sought to identify topics with high salience at points of time in the study period. Specifically, in order to determine frame elements to include in the analysis, it was

necessary to determine a threshold proportionality to distinguish highly salient topics at a given time from less salient topics (as described in Chapter Two, topics with higher proportions in the corpus are considered more salient). In order to determine this threshold, I examined the proportional coverage for all 38 defined topics and found that most of the topics remained below a proportion of 0.03 across time but that a smaller subset of the topics exhibited proportional coverage above 0.03 for certain periods of time. Therefore, I designated 0.03 proportionality as the threshold for topics to be considered relatively high proportions in the corpus and therefore salient frame elements. Across the 20-year period, 19 of the 38 topics had high salience at some point in time using this threshold. By analyzing how clusters of these 19 topics vary over time, I develop an understanding of the structure of the issue space and can assess the extent to which the framing of testing evolved over time.

In this section, I highlight and describe the proportionality of several of the salient topics in the dataset that exhibit particularly interesting patterns of coverage. Coverage of the topic about the No Child Left Behind Act was highly salient (greater than 0.03 proportion) from mid-2003 to 2009. As can be seen in Figure 4.5, coverage of this topic began to rise around 2001, when the legislation to reauthorize the education law was introduced. Coverage continued to rise until 2007 and then fell precipitously from 2007 through 2010 before leveling off somewhat through 2015. The peak in coverage from 2005 through 2006 may be due to a convergence of several related issues regarding NCLB and testing. First, there may have been increased interest in how states were responding to the testing mandate portion of the law, which required states to implement annual testing in reading and math by the 2005–2006 academic year. Additionally, the law was due for reauthorization in 2007, which appears to have sparked additional interest in

assessing the impact of the law. For example, an article from the *New York Times* in November 2006 with the headline "Schools Slow in Closing Gaps Between Races" begins as follows:

> When President Bush signed his sweeping education law a year into his presidency, it set 2014 as the deadline by which schools were to close the test-score gaps between minority and white students that have persisted since standardized testing began. Now, as Congress prepares to consider reauthorizing the law next year, researchers and a half-dozen recent studies, including three issued last week, are reporting little progress toward that goal […] (Dillon, 2006b).

Tests, as this excerpt illustrates, can be used to highlight disparity in achievement between subgroups of students. This was one of the purposes of the testing requirement of the law: to draw attention to achievement gaps and hold schools accountable for closing those gaps.

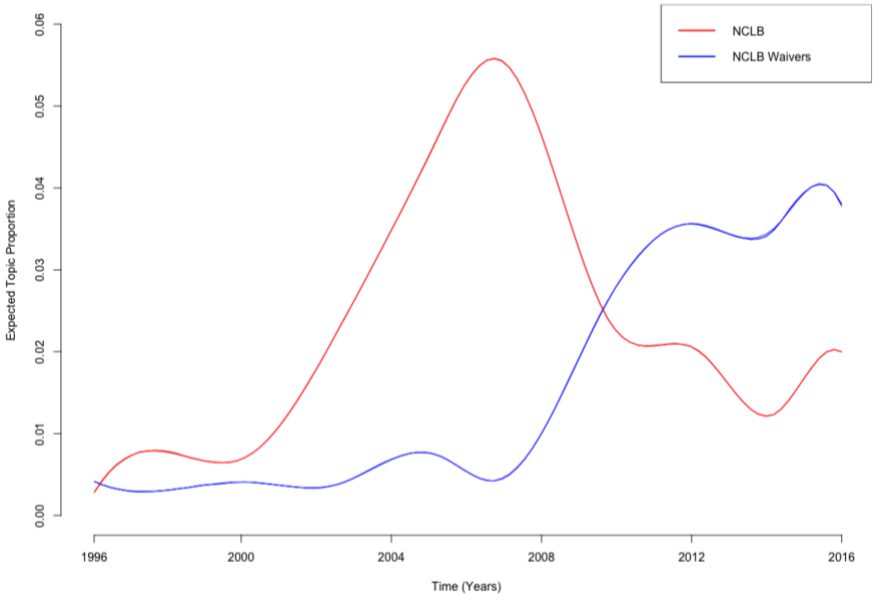
Another factor contributing to the rise in coverage of NCLB was the state-level response to the law, including issues of setting standards and measuring adequate yearly progress. For example, an article from November 2005 in *Education Week* titled "Shifts in State Systems for Gauging AYP Seen as Impeding Analysis" described how states had developed varying measures of progress, which complicated attempts to make comparisons across states:

> Determining whether schools and districts are making adequate yearly progress under the federal No Child Left Behind Act "has evolved into 50 intricate formulas that vary greatly from state to state," according to a recent report by the Center on Education Policy. The report from the Washington-based policy group tracks changes to state accountability plans approved by the U.S. Department of Education in 2004 and 2005, based on decision letters posted on the department's Web site […] (Olson, 2005).

These issues likely contributed to the spike in coverage of NCLB and testing that peaked between 2006 and 2007.

A related topic to NCLB was coverage of NCLB waivers, which is also depicted in Figure 4.4. Trends in the coverage of these two topics suggest that as discussions of a reprieve from aspects of NCLB among policymakers became more prevalent in 2009 and 2010 and waivers were formally introduced by the federal government in 2011, coverage of these waivers largely replaced coverage of NCLB itself. Implications of this shift in coverage will be discussed further in the section on frames.

Figure 4.5. Coverage of NCLB and NCLB Waivers, 1996 – 2015

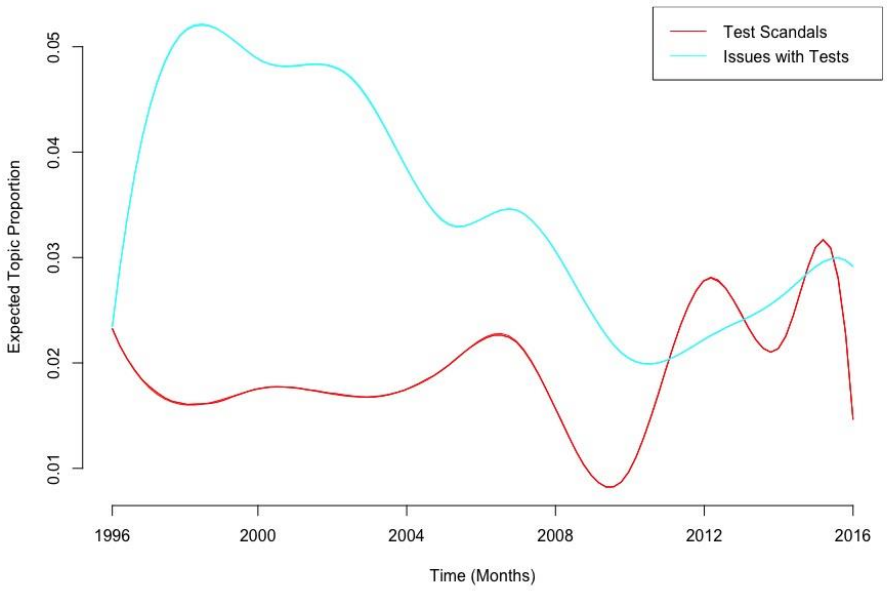


Another set of topics that appeared to be related were the Issues with Testing topic and the Testing Scandals topic. Patterns in coverage suggest that for approximately the first ten years of the dataset, proportional coverage of the two topics exhibited very different patterns. These two topics were not moving together over time and the Issues with Testing topic was more prevalent. However, starting around 2006 through 2015, the two topics moved in concert and coverage of issues with testing rose along with increased coverage of testing scandals. As shown in Figure 4.6, after a drop in coverage of testing scandals in 2008, this topic proportion rose

substantially beginning in 2009 and remained relatively high through the remainder of the dataset, reaching a high in 2015. Along with this rise, the Issues with Testing topic, which had similarly dropped to a 20-year low in coverage around 2008, showed an increase in coverage starting in 2010 through 2015. This pattern suggests that as testing scandals gained attention in the media, discussions of issues with testing, which had fallen off around 2007, became more prominent again. An example article from *Education Week* in the wake of the Atlanta testing scandal, for example, reported on the varying responses to the cheating issue, including calls to reduce accountability pressures by eliminating high-stakes testing:

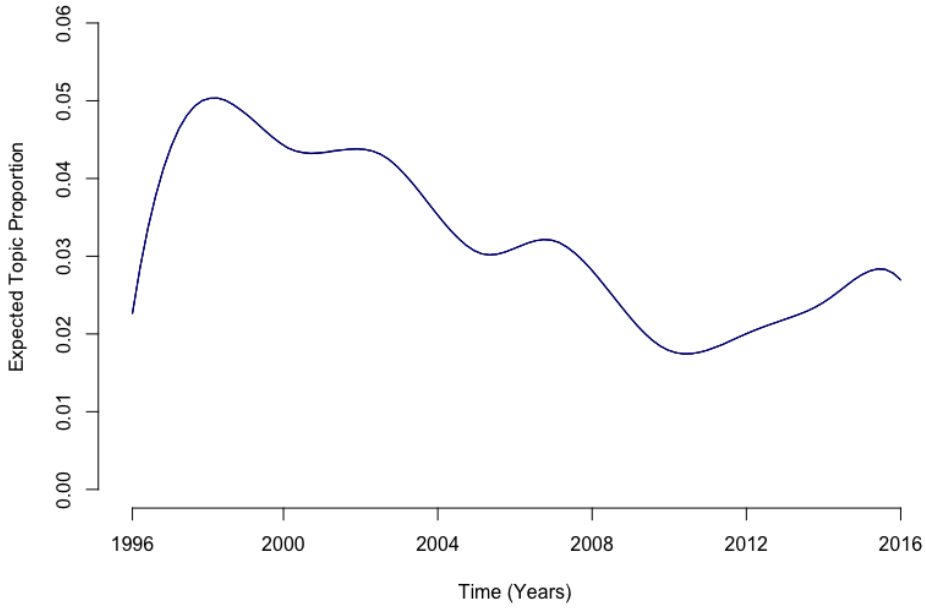
> The allegations of systematic test alteration by teachers and principals in Atlanta, along with recent accusations of cheating in Baltimore, the District of Columbia, Philadelphia and other districts, have highlighted a split between those arguing for improved test management and security and those who ask if it's better to scrap high-stakes testing altogether (Samuels, 2011).

Figure 4.6. Coverage of Issues with Testing and Testing Scandals



The one explicitly negative frame element that was salient in the dataset was coverage of issues with tests. This frame element was highly salient throughout the years of coverage with the exception of the period of years from 2009 to 2013. This suggests that the media have consistently covered issues with tests to a relatively high degree since the mid-1990s. Coverage of issues with testing was particularly high from 1998 to 2004 (see Figure 4.7). Interestingly, the trends depicted in Figure 4.7 suggest that coverage of the Issues with Testing topic was highest during the early years of the study and dropped in proportion across the 20-year period. This pattern suggests that discussions of potential problems and misuses of testing received media coverage prior to the passage of NCLB and its annual testing requirements, rather than being a result of the increased testing requirements under the law.

Figure 4.7. Coverage of the Issues with Testing Topic, 1996 – 2015

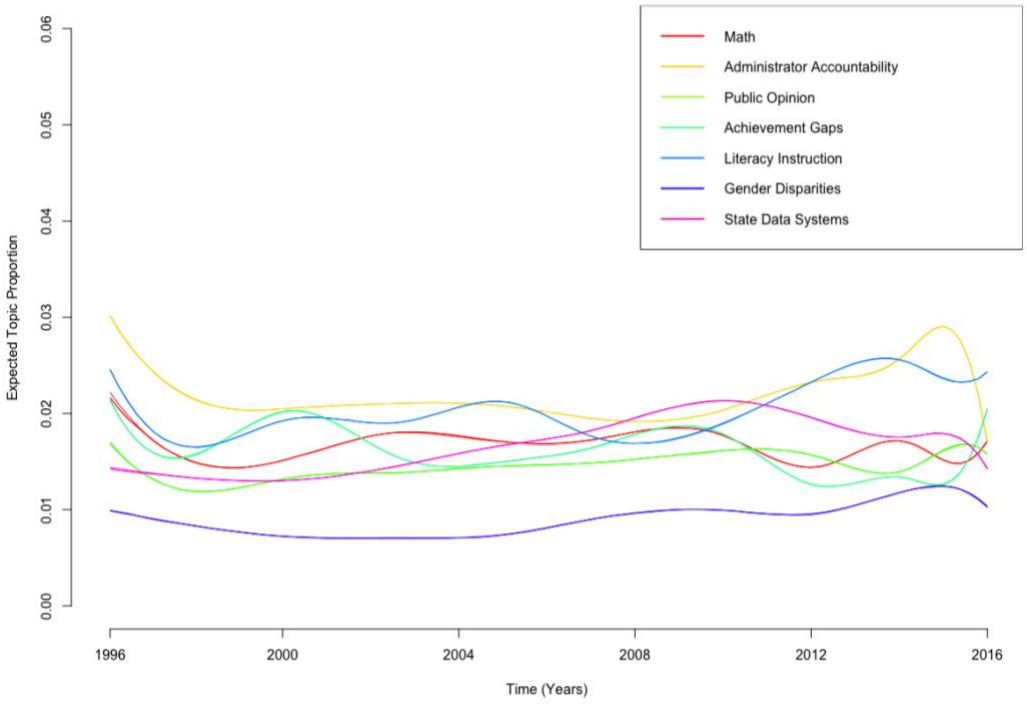


Topics with Minor Variation in Coverage

Although many of the topics in the topic model had substantial variation in proportionality across time (see Figure 4.3) and 19 of the topics exhibited relatively high salience at some point in time, seven of the 38 topics exhibited very little variation in coverage over the

twenty-year period. These topics included coverage of math, administrator accountability, public opinion, achievement gaps, literacy instruction, gender disparities, and state data systems. In addition to relatively consistent levels of coverage across time, all of these topics had relatively low salience (remaining below the 3 percent threshold I established for determining high salience of topics). This is illustrated in Figure 4.8. Because these topics had consistently low salience, they were not considered frame elements as described in the conceptual model.

Figure 4.8. Flat Topics with Low Salience Across Time



Among those topics identified as having relatively low salience across time, several stood out as unexpected, including achievement gaps and math. Given that achievement gaps between subgroups of students were a substantial focus of NCLB, both in terms of framing the need for the law and also in terms of defining adequate yearly progress, I expected this topic to exhibit some spikes in coverage, particularly around the time of passage of NCLB and during the years immediately following passage. One possible explanation is that there is some overlap between this topic and other topics in the model that are related, such as trends in test scores, NAEP, math

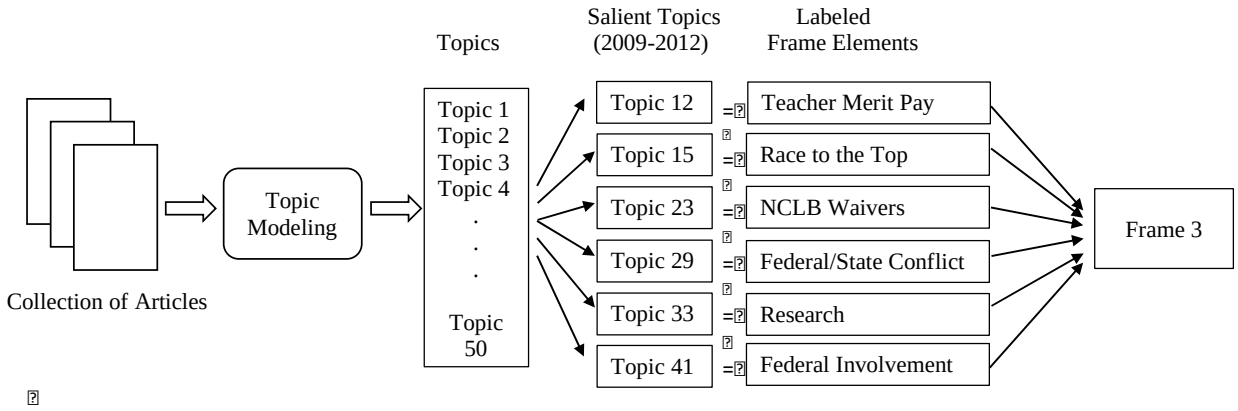
achievement, or writing achievement. For example, NAEP test score data is often used to identify achievement gaps (Porter, n.d.). Similarly, the emergence of new topics that become salient may serve as a proxy for this topic. The lack of coverage of the math topic as it relates to testing was also unexpected, particularly given that math is one of the subjects with mandated annual testing. Again, this topic may have had reduced salience in the dataset because coverage of math overlapped with other topics, particularly given the importance of math test scores in discussions about many other issues related to testing.

Frames

As discussed in the description of the conceptual framework in Chapter Two, topics from the topic model are conceptualized as frame elements. Specifically, I identified frame elements as those topics with relatively high salience (above 0.03 proportionality). Once these frame elements were identified, I was able to track how these elements cohere and form frames. This is illustrated in Figure 4.9 below, which represents the process of determining the frame elements for the period of years from 2009 to 2012 (Frame 3) from the topic model output.

By utilizing topic modeling to identify topics, then determining which topics are highly salient (and therefore are frame elements), and then assessing how these frame elements cluster together, it is possible to gain insight into the media framing of the issue and whether and how this framing evolved over time. The process of analyzing frames, rather than solely assessing individual topics, provides insight into the broader ways in which the media may be shaping public and policy actors' perceptions of the issue of testing.

Figure 4.9. Identifying Frame Elements for a Frame Using Topic Modeling



Combinations of highly salient frame elements create frames. The power of frames can then be measured using frame element salience, resonance (the number of frame elements that constitute a frame), and persistence (the length of time that frame elements cluster together to create a frame). Using the results of the topic modeling and mapping of highly salient frame elements, I identified four major frames across the 20-year period.

For each frame, I analyzed the frame elements to determine the extent to which each element was associated with either positive, neutral, or negative aspects of testing. For example, the Research topic was labeled as positive because articles highly associated with this element were largely about how test scores were being used in research studies of education issues. Although this is not an overt message in favor of testing, it privileges tests as the means of assessing effectiveness of educational interventions and improved test scores as a primary outcome of interest. In an illustrative article from *Education Week* reporting on a federal study of supplemental reading programs, the programs were deemed ineffective at improving reading comprehension based on test scores (Gewertz, 2010). Again, this coverage of testing was not explicitly positive in terms of advocating for testing, but the underlying message of articles citing research that used test scores as the measure of improvement is positive. Conversely, the

Federal/State Conflict frame element was labeled as negative because articles associated with this element were largely about state pushback to federal involvement in education, specifically testing mandates.

This process allowed me to determine the extent to which the frames change and also whether and how the frames evolve in terms of overall tone. Table 4.4 provides details for each of the four frames. With the exception of Frame 4, all of the frames had a combination of positive, neutral, and negative frame elements. Frame 4 consisted only of neutral and negative frame elements and did not have any positive elements.

New frames were identified as the result of changes in the frame elements (topics), which consisted of both the addition of newly salient frame elements and the subtraction of frame elements that were no longer salient. Several of the frame elements were consistent across multiple changes in framing. For example, the issue of school performance ratings, which was about policies and processes to give schools grades or other ratings that are largely based on student test scores, was a salient element across the first two frames. Other frame elements were only salient for relatively brief periods of time. The testing scandals topic, for example, was only salient over a period of approximately ten months in 2015, at the end of the study period. None of the four frames were composed of entirely unique frame elements, which was expected given that only 19 of the 38 total topics exhibited relatively high salience at some point across the 20-year period of the study.

Two frame elements (Teacher Merit Pay and Federal Involvement) were consistently present across the study period. I labeled both of these elements as neutral because neither element was strongly associated with either positive or negative aspects of testing. For example, although teacher merit pay is a controversial issue in education, the specific connection with

testing in media coverage was not predominantly positive or negative. To illustrate, an article on the union's stance on a teacher merit pay plan in New York City from the *New York Times* in 2001 mentioned testing only within the context of providing a measure of student performance that can be used to provide bonuses:

> The union, the United Federation of Teachers, indicated over the last week that it would seriously consider a plan that awards all of a school's teachers and other staff members merit pay, perhaps as a one-time bonus, when the school's students show overall improvement on various measures, including standardized tests (Greenhouse, 2001).

As another example, an article from 2007 in *Education Week* on a teacher merit pay plan in Texas mentioned that bonuses would be given to teachers "for achievement in improving test scores and other signs of student progress." These examples serve to illustrate that testing coverage as it related to the issue of teacher merit pay was predominantly neutral. The same is true for coverage of the issue of federal involvement in education (as well as the other frame elements that I identified and labeled as neutral).

Although the two frame elements described above were consistent across the study period, each of the four frames exhibited a substantial shift in frame elements from the previous frame (thus identifying it as a new frame), while also maintaining some of the frame elements from the previous frame. For example, in the shift from Frame 1 to Frame 2, four frame elements dropped in coverage and therefore were not elements of Frame 2 (Regents Exams, State-level Elections, $4^{th}/8^{th}$ Grade Tests, and NAEP) and three frame elements that were not part of Frame 1 increased in coverage and became part of Frame 2 (NCLB, Federal/State Conflict, and Research). In addition to these changes, seven frame elements remained the same in the

shift from Frame 1 to Frame 2. The shifts from Frame 2 to Frame 3 and from Frame 3 to Frame 4 exhibited similarly substantial changes in frame elements (see Table 4.3).

The composition of each frame and the specific combination of positive, negative, and neutral frame elements in each are discussed in greater detail below.

Table 4.3. Frames and Salient Frame Elements by Tone in Testing Coverage

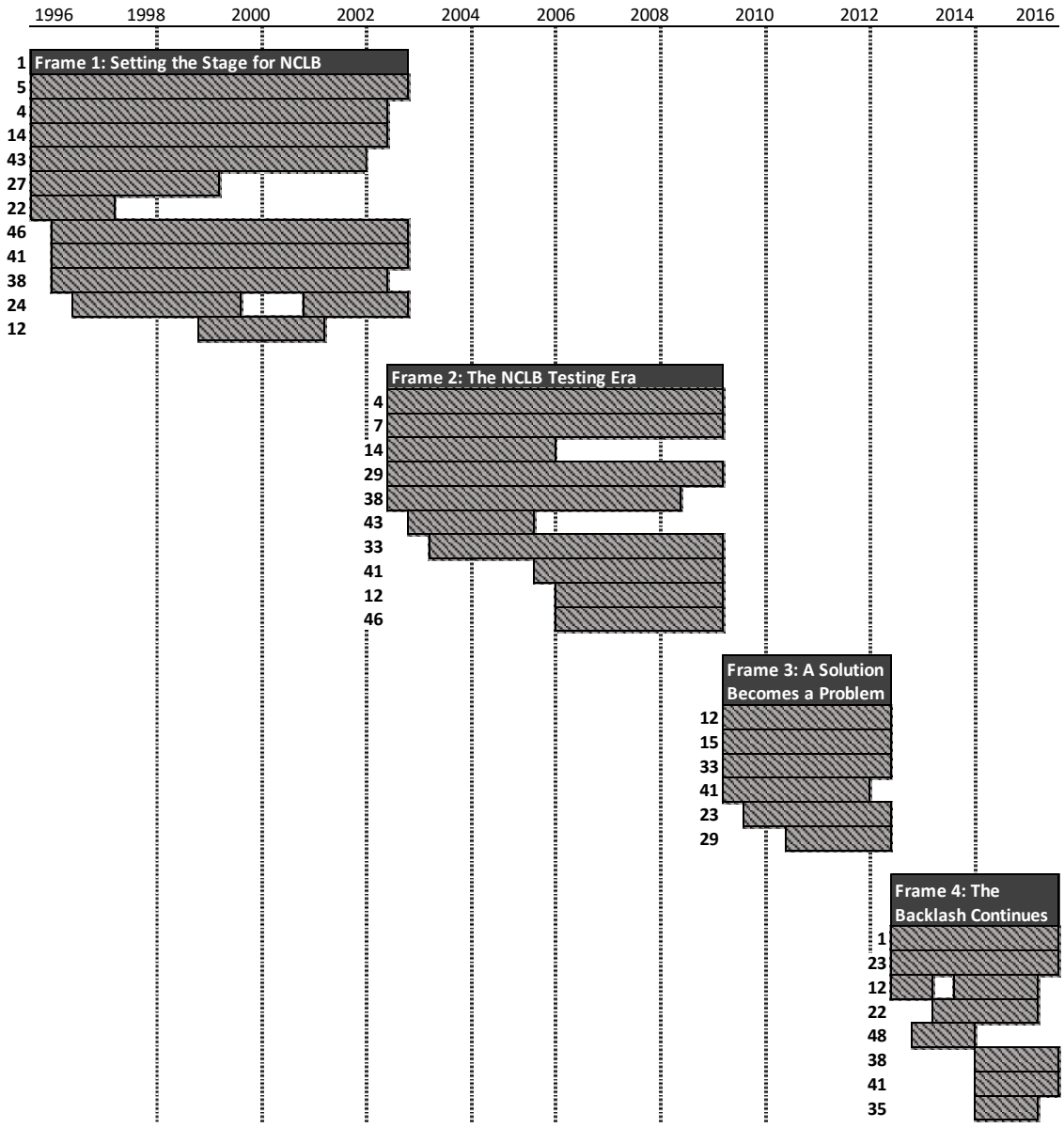
FRAME 1 1996-2002	FRAME 2 2002-2009	FRAME 3 2009-2012	FRAME 4 2013-2016
School Performance Ratings	School Performance Ratings	Race to the Top	Testing Scandals
Regents Exams	NCLB	NCLB Waivers	NCLB Waivers
Teacher Merit Pay	Teacher Merit Pay	Teacher Merit Pay	Teacher Merit Pay
High School Exit Exams	Federal/State Conflict	Federal/State Conflict	State-level Elections
State-level Elections	Research	Research	Common Core Controversy
Issues with Tests	Issues with Tests	Federal Involvement	Issues with Tests
Federal Involvement	Federal Involvement		Federal Involvement
4th/8th Grade Tests	High School Exit Exams		Technology
NAEP	Trends in Test Scores		
Trends in Test Scores	Education Finance		
Education Finance			

Note: Frame elements shaded in green are positively associated with testing. Frame elements shaded in red are negatively associated with testing. All other frame elements are considered neutral.

Figure 4.10 illustrates the four frames that cover the period of time from 1996 through 2015. The first two frames each covered an extensive period of time. Frame 1 lasted for

approximately six years from 1996 to 2002 and Frame 2 lasted for approximately seven years from 2002 to 2009. The last two frames covered approximately the same length of time (about 40 months or over 3 years of coverage). It is important to note that the changes in framing identified in the current study occurred gradually. Although there were some instances where specific policy developments led to a dramatic increase or decrease in the salience of a frame element, overall the shifts to a new frame were incremental. For the purpose of analysis and to clearly represent the shifts to new frames, Figure 4.10 clearly demarcates each frame. It is important to remember, however, that the shift occurs gradually rather than abruptly and some frame elements remain constant across time even as the underlying frame shifts to a new understanding of the issue.

Figure 4.10 Four Frames of Media Coverage of Testing, 1996 – 2015

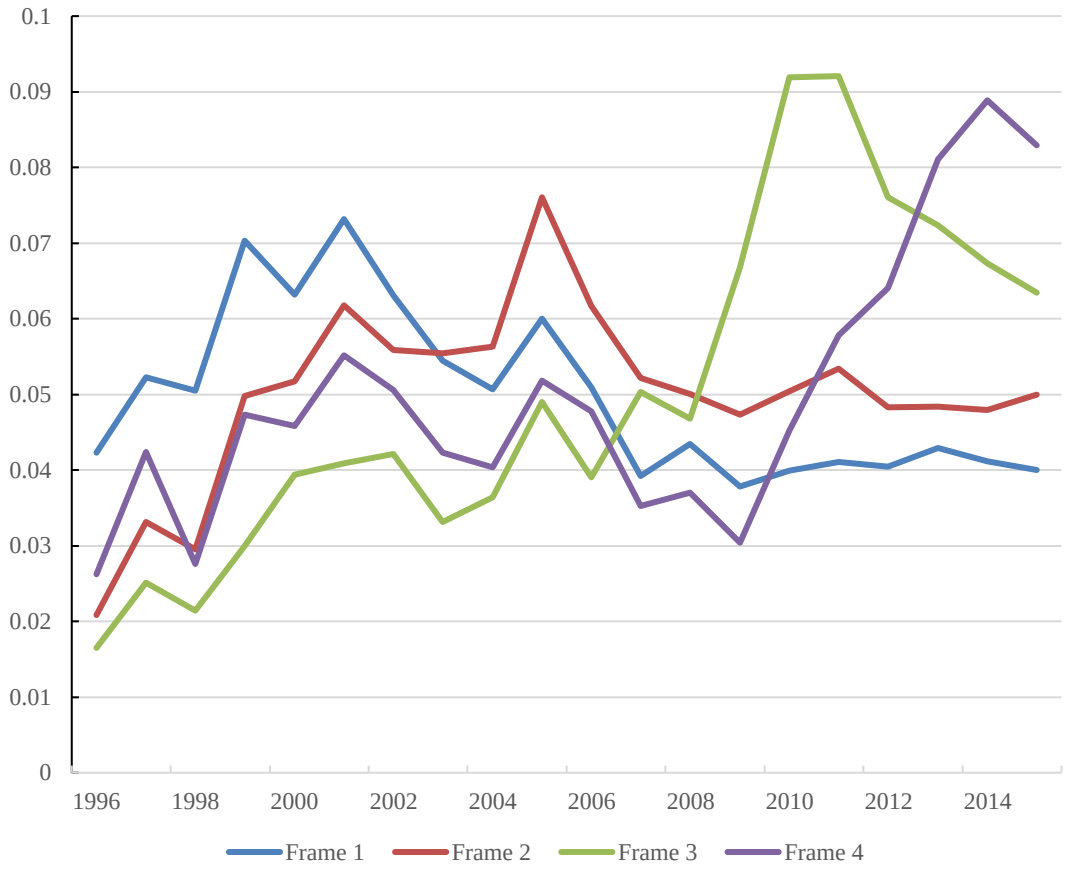


Note: Numbers listed before each frame element are the topic numbers. See the Appendix for a full list of topics by number.

The shifts between frames can also be assessed by plotting each frame as a variable over time. These variables are created using counts of articles for each frame over time and are then plotted as percentages. Figure 4.11 shows the four frames plotted across the 20-year study period. As can be seen in the figure, there are clear shifts in the dominant frame over time. In

particular, Frame 3 and Frame 4 show dramatic rises (Frame 3 started to rise in proportion around 2008 and Frame 4 started around 2010).

Figure 4.11 Four Frames Plotted as Variables Across Time



Interpreting Frames

Because each frame is composed of multiple frame elements, the frames are not easily interpretable in terms of being either clearly positive framing of testing or negative framing of testing. However, as I will explain below, the findings from the topic modeling do indicate a shift in framing to more negative and controversial aspects of testing in the last years of the study. In the following sections, I describe my analysis and interpretation of the frame elements of each frame and utilize the concepts of salience, resonance, and persistence described in Chapter Two to measure the power of each frame. Salience is determined by the proportional

coverage of the element compared to other elements (with 3 percent and above indicating high salience), resonance is determined by the number of frame elements that constitute a frame (more frame elements indicate a more resonant frame), and persistence is determined by the length of time that a frame exists (longer frames have higher persistence).

Frames 1 and 2: Positive Coverage and the Rise of NCLB

Frame 1: Setting the Stage for NCLB. The first frame covered the time period from 1996 to 2002, which are the six years leading up to the passage of NCLB. This frame contained the highest number of salient frame elements (11 frame elements) and also lasted for the longest period of time, making it a particularly salient, persistent, and resonant (and therefore powerful) frame. This frame included a higher proportion of frame elements that are associated with positive characteristics of testing or topics associated with arguments made by proponents of testing. For example, proponents of testing have noted that test scores provide important data for calculating school performance ratings, which helps parents and students make informed decisions about schooling and provides comparable information across schools. An article from *Education Week* published in 2000 about the unveiling of California's ranking system, which was based on standardized test scores, for example, stated, "Schools' ability to compare themselves with similar schools is one of the most important features of the state [ranking system]" (Sandham, 2000). Similarly, high coverage of NAEP during the pre-NCLB years suggests a focus on tests as a way to measure schools, districts, and states and as a tool to provide meaningful, comparable data on performance. For example, an article in the *New York Times* from 1997 stated, "In a generally encouraging sign of progress in American education, the latest report from the only national assessment of educational achievement [NAEP] shows steady and significant progress in mathematics test results by the nation's fourth-, eighth-, and 12th-

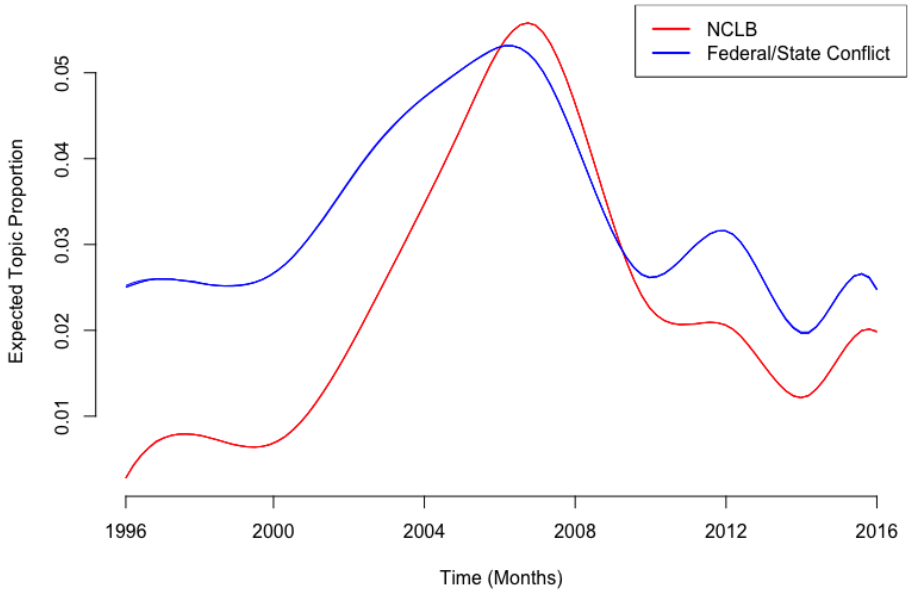
grade students" (Applebome, 1997). Implicit in this statement is the idea that test scores are valid measures of educational progress for the nation. Frame 1 included six positive frame elements, one negative element, and four neutral elements. The relatively high level of attention to positive aspects of testing during this first framing may have helped set the stage for the subsequent passage of NCLB, with its focus on annual testing as a core component of the law.

The first (and second) frame included a high proportion of coverage of high school exit exams and Regents exams (Regent exams are required in New York State for high school graduation), which suggests a focus during the first ten years of coverage on the use of these tests to increase rigor in schools or as a kind of credentialing function to demonstrate that students have learned appropriate content. Proponents of these types of exit exams have proposed them as a solution to the problem of social promotion. During the Clinton administration, in particular, social promotion was a controversial topic and more testing was frequently proposed as a solution to the problem of students being promoted to subsequent grades without mastering content (Wachen, 2014). High school exit exams have existed for decades but became particularly popular among state policymakers in the late 1990s to early 2000s (Education Commission of the States, n.d.). However, some states started to move away from requiring students to take these tests in later years—nine states have recently ended exit exams as a graduation requirement and several other states have reduced the number or weight of these tests ("Graduation test update," 2017).

Frame 2: The NCLB Testing Era. Frame 2 consisted of 10 frame elements and contained four positive elements, two negative elements, and four neutral elements, suggesting a less positive, more mixed framing of testing. Seven of the frame elements stayed consistent across Frames 1 and 2: School Performance Ratings, Teacher Merit Pay, High School Exit

Exams, Issues with Tests, Federal Involvement, Trends in Test Scores, and Education Finance. However, there were also three new frame elements and four elements that were no longer salient. During the period of time covered by this second framing of testing (2002—2009), coverage of NCLB and tensions between the federal government and states became two highly salient issues. The NCLB frame element was coded as neutral because coverage of this topic was neither overly positive nor negative in relation to testing. The Federal/State Conflict frame element was coded as negative due to its emphasis on disagreements between the states and the federal government, particularly regarding state pushback to the testing mandate. These two frame elements appear to be connected. Figure 4.12 illustrates that these two elements are moving together across time. The increased federal involvement in education that accompanied NCLB (or at least the perception that NCLB was a substantial and unprecedented push by the federal government) created tension between the U.S. Department of Education and state-level policymakers. Connecticut's legal challenge to NCLB is one example of that tension playing out in the states. The patterns in the proportional coverage suggest that an increasingly dissonant relationship between the federal government and the states with regard to education policy and testing (and media coverage of this development) was occurring prior to the passage of NCLB and that coverage of the relationship continued to increase through the years following passage until peaking in 2006. As coverage of NCLB began to drop in 2007 so too did coverage of the conflict between the federal government and the states.

Figure 4.12. Coverage of NCLB and Federal/State Conflict



Moreover, the rise of coverage of NCLB coincided with a rise in coverage of research utilizing test score data. NCLB's annual testing requirements in grade 3-8 and once in high school resulted in an increase in available data on student and school performance. Researchers were able to access this newly available data for studies of student performance. Because annual testing was part of the law and required in all states (states could, in fact, choose not to comply with the law, but they risked losing substantial federal funding), these new data were also more extensive and allowed for more fine-grained analyses. An article published in *Education Week* in 2008 that was associated with the Research topic, for example, noted how NCLB data was an improvement over previously available federal data:

> The amount and quality of data available today represent a dramatic improvement over what was available in the so-called "wall chart," a state-by-state compilation of resource inputs, performance outcomes, and population characteristics that the Education Department published for six years, starting in 1984 under Secretary Terrel H. Bell (Hoff, 2008).

Frames 3 and 4: The Rising Coverage of Problems with Testing

Frame 3: The Testing Solution Becomes the Testing Problem. Frames 3 and 4 differed from Frames 1 and 2 in terms of persistence, lasting for approximately three years each. With only six frame elements, Frame 3 had the lowest resonance of the four frames. Unlike previous frames, Frame 3, which includes the period from approximately 2009 to 2012, did not contain a high proportion of frame elements associated with positive aspects of testing, signifying an important shift in coverage. Coverage of three positive frame elements (school performance ratings, high school exit exams, and trends in test scores) dropped during this period. The only positive frame element in Frame 3 was the Research topic. Also during this time period, a high level of coverage of NCLB was largely replaced by coverage of NCLB waivers, which is reflected in Figure 4.5. This change suggests an increased focus on the aspects of NCLB that were perceived as negative and that led to the federal government's recognition that some of the mandates of the law may have been misguided.

As discussions of a possible reprieve began to take shape in the years leading up to the announcement of waivers in 2011, coverage of NCLB started to highlight negative aspects of the law. For example, in an article titled "New Tack on NCLB: Regulatory Relief" from 2010 in *Education Week*, the law is described as "onerous" and "inflexible and intrusive." Later, in a 2013 press release announcing that states could reapply for waivers, U.S. Secretary of Education Arne Duncan stated that NCLB was "outmoded and constrains state and district efforts" (U.S. Department of Education, 2013). Although the waivers were not specifically designed to address testing issues, a core component of NCLB was the annual testing mandate which was directly tied to accountability and sanctions. Efforts during this period of time to offer waivers and attempts to reform the law through reauthorization reflect a growing belief among policymakers

that NCLB and its heavy emphasis on high-stakes testing was not beneficial for schools or students.

Another distinguishing feature of Frame 3 was the high level of coverage of Race to the Top. Introduced by the federal government in 2009, Congress appropriated over $4 billion for the program as part of the American Recovery and Reinvestment Act, and the program was funded through September 2015. However, coverage of this topic dropped off and was no longer salient by 2013 (several years before the grant program ended). This suggests that when Race to the Top was first introduced and as states began to respond, media outlets were more interested in closely following developments than in later years as the program matured. The Race to the Top frame element was coded as neutral in the analysis because there was not a predominant association between the grant program and arguments in support of or against testing.

However, in order to be eligible for Race to the Top funds, states were required to submit plans for developing systems to evaluate the effectiveness of teachers. These evaluation systems were based largely on student test scores, which was a controversial aspect of the program. Teachers' unions, for example, pushed back against the emphasis on using tests for evaluating teachers. For example, a 2009 article from *Education Week* covered the National Education Association's statement in response to the announcement of Race to the Top and noted, "Among other areas, the NEA says it cannot support the fund's endorsement of using test scores for evaluating teachers" (Sawchuk, 2009). This conflict over aspects of the grant program suggests that although overall coverage of Race to the Top and testing was not negative, there was a concern among some constituents over the growing use of test scores for other purposes such as evaluating teachers.

This pushback to the expansion of the use of tests, coupled with the negative frame elements (NCLB Waivers and Federal/State Conflict), suggests that Frame 3 was a turning point in media coverage of testing, in which the predominant framing of testing shifted from positive to negative. This shift continued a trend that was present in the evolution from the first, highly positive framing of testing (during the pre-NCLB years) to a less positive second frame, as coverage of positive aspects of testing decreased and was largely replaced by an increase in coverage of topics negatively associated with testing. Frame 4, discussed next, further solidified this shift.

Frame 4: The Backlash Continues. Frame 4 persisted until the end of the study period (approximately 3 years from 2013 through 2015). In terms of resonance, this frame included eight elements, making it more resonant than Frame 3. However, several frame elements were present for relatively brief periods of time, which may be an indication that this frame was still developing when the current study was completed. In addition to increased media coverage of NCLB waivers during the period of time covered by the last two frames, several other frame elements that became salient during the period of time from 2013 through 2015 indicate a substantial shift toward a negative framing of the testing debate, including coverage of testing scandals and controversy about the Common Core State Standards. These two new frame elements are discussed below.

***Testing scandals*.** The model indicates that there was some relatively minor coverage of testing scandals in states and districts around the country in the late 1990s and early 2000s (for example, in late 1999, both *Education Week* and the *New York Times* covered a controversial claim of widespread instances of teachers helping students on tests in New York City schools) but the dramatic increase in the salience of this frame element in the 2013-2015 period was

primarily driven by the high level of publicity of the Atlanta cheating scandal and the subsequent trial and conviction of administrators and teachers. Additionally, the Atlanta cheating scandal prompted public discourse about potentially perverse consequences of testing when used for accountability purposes. For example, an *Education Week* article from 2015 about the conviction of several of the educators involved in the Atlanta scandal included a discussion about these consequences, quoting several scholars engaged in the debate and noting:

> The widespread nature of the alleged cheating in Atlanta and other districts in recent years has helped fuel a national debate about high-stakes standardized tests in schools: the frequency of those tests, and the ease by which the scores could be manipulated by a few (Mitchell, 2015).

Common Core controversy. By the end of 2010, 43 states and the District of Columbia had adopted the Common Core (Rothman, 2011). In addition to concern among some state policymakers and education stakeholders about the use of the common standards themselves, the standards-aligned tests designed by the two testing consortia were also controversial. The two consortia (PARCC and Smarter Balanced) began administering tests aligned to the standards in the 2014—2015 school year. However, states began to back out of the consortia in the past few years. State membership in PARCC, for example, dropped from 24 to 12 by the fall of 2015 (Camera, 2015). Findings from the study indicate that media coverage of the controversies about the standards and the standards-aligned tests began to increase dramatically in 2012—2013 and peaked in 2014 (see Figure 4.4). An article from *Education Week* in 2014, for example, stated, "The formal structures that buttress the standards, and the related tests from two federally funded consortia, have eroded somewhat, as states reconsider their adoptions of the standards and reject the common tests" (Ujifusa, 2014b). It remains unclear whether media coverage influenced state

actions or state actions influenced media coverage, but it may be that these developments influenced each other and both played a role in the growing backlash to the Common Core.

These newly salient and negative frame elements were accompanied by an absence of elements that were more positively associated with testing, in contrast to earlier frames (see Table 4.4). For example, Frame 1 included salient frame elements about high school exit exams, NAEP, school performance ratings, and trends in test scores. None of these elements were salient during the period of time covered by Frame 4. Taken together, this change indicates a shift from coverage of topics and issues related to testing that have more positive associations during the first two frames to coverage of issues that have a more negative association with testing.

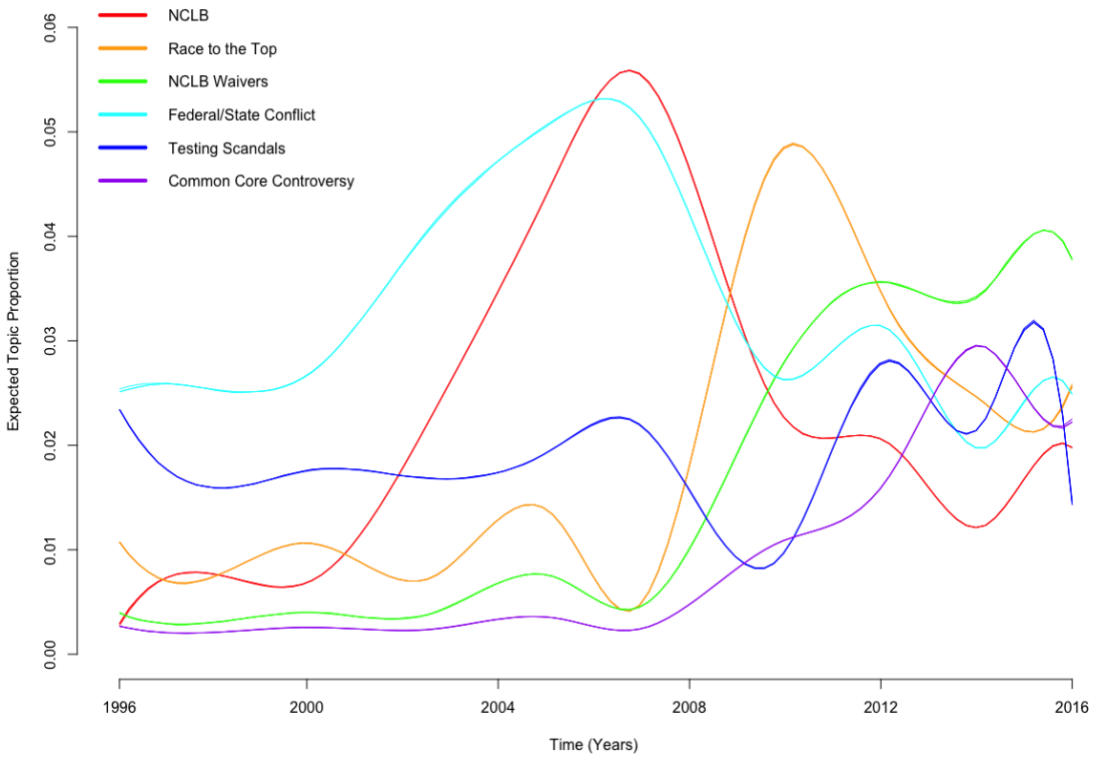
NCLB and Negative Coverage of Testing

The passage and subsequent controversy surrounding NCLB, and the law's testing mandates specifically, may have influenced the more negative framing of the issue of testing in later years. As shown in Figure 4.13, in the years after the rise of coverage of NCLB in the early 2000s, several topics increased in coverage that were more negative, including the Federal/State Conflict topic, Testing Scandals topic, and Common Core Controversy topic. At the time of passage of NCLB, politicians heralded NCLB's testing mandate as an important tool for holding schools accountable for improved student achievement. In an illustrative example, U.S. Secretary of Education Rod Paige remarked on the importance of testing in the law:

> Some worry that we have placed the emphasis on tests, not teaching. I am surprised by the debate about the need for tests. How else can we measure if students are learning? […] Testing allows us to highlight the students who most need our help so we can give them the help they need" (Paige, 2003.)

Placing this level of attention and importance on testing through NCLB may have ushered in a period of scrutiny of this purpose of testing and resulted in increased media coverage of possible unintended negative consequences.

Figure 4.13. Coverage of NCLB, Race to the Top, and Topics Related to Negative Aspects of Testing



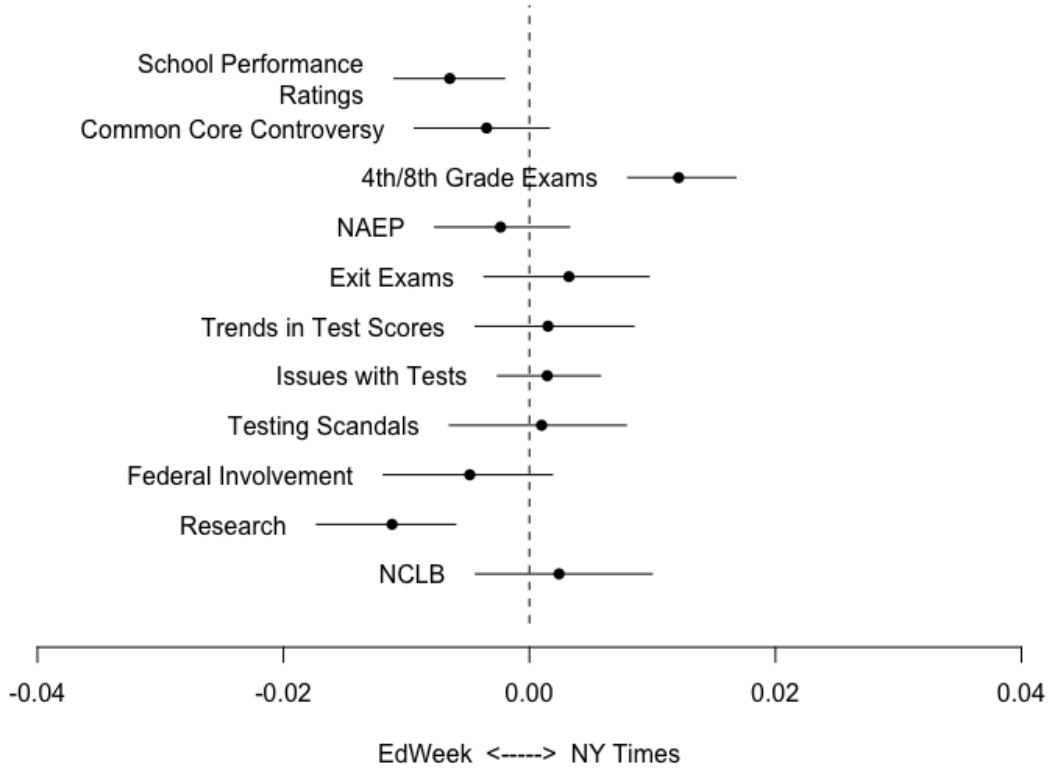
Frame Elements by Source

One way to examine differences between topic coverage in *Education Week* and the *New York Times* is to graphically depict the mean difference in topic proportions for the two publications. This is done using the *plot.estimateEffect* function in the *stm* package in R with the "difference" method specified.

As shown in Figure 4.14, which plots the mean differences for both positive and negative frame elements, the frame elements generally do not show relatively large differences in proportions between *Education Week* and the *New York Times*. When accounting for confidence

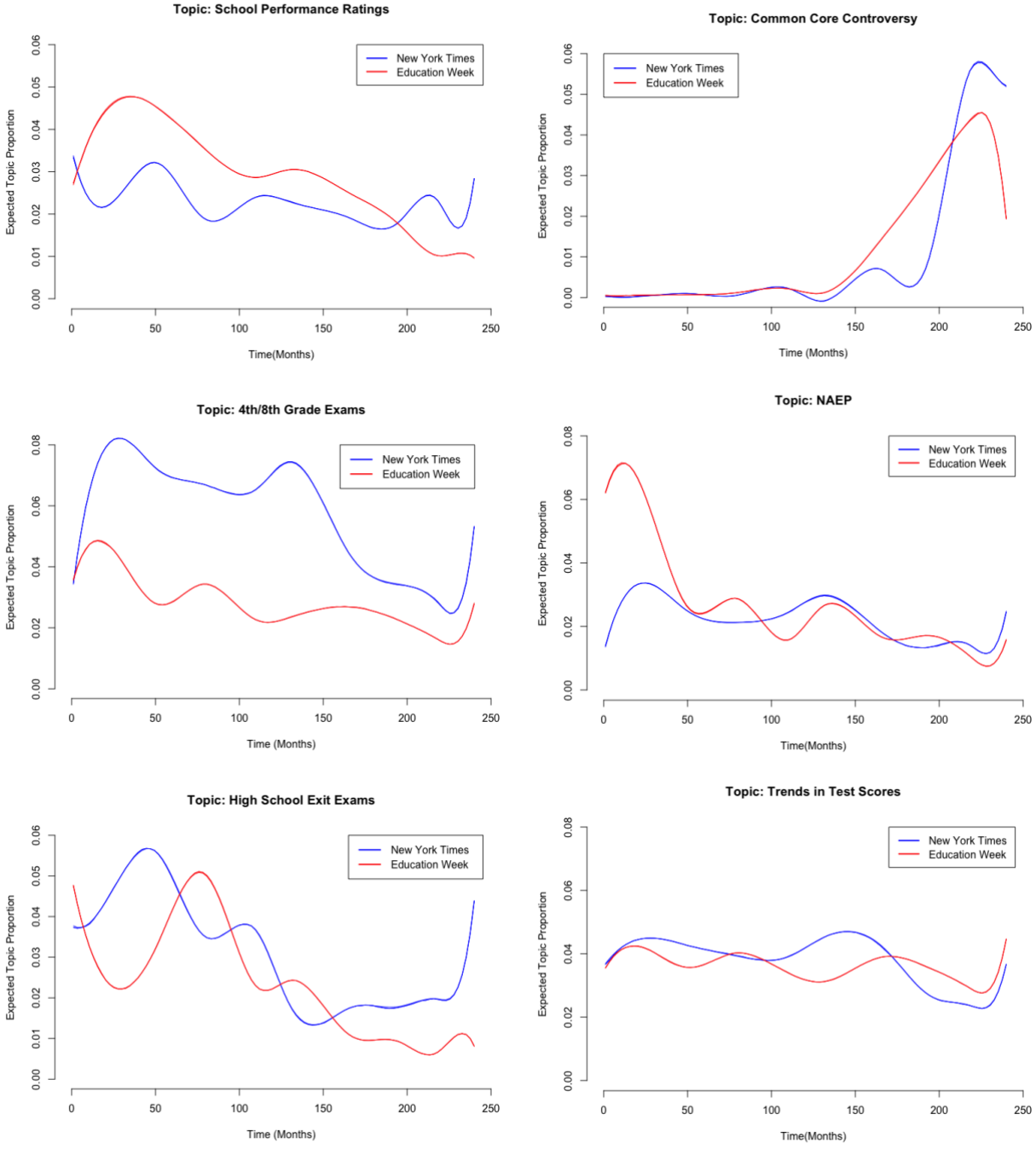
intervals, only three of the frame elements had significant differences between the two publications: School Performance Ratings, 4th/8th Grade Exams, and Research. The School Performance Ratings and Research topics exhibited higher proportions of coverage in *Education Week* and the 4th/8th Grade Exams topic had higher coverage in the *New York Times*. Differences in these three topics are discussed below.

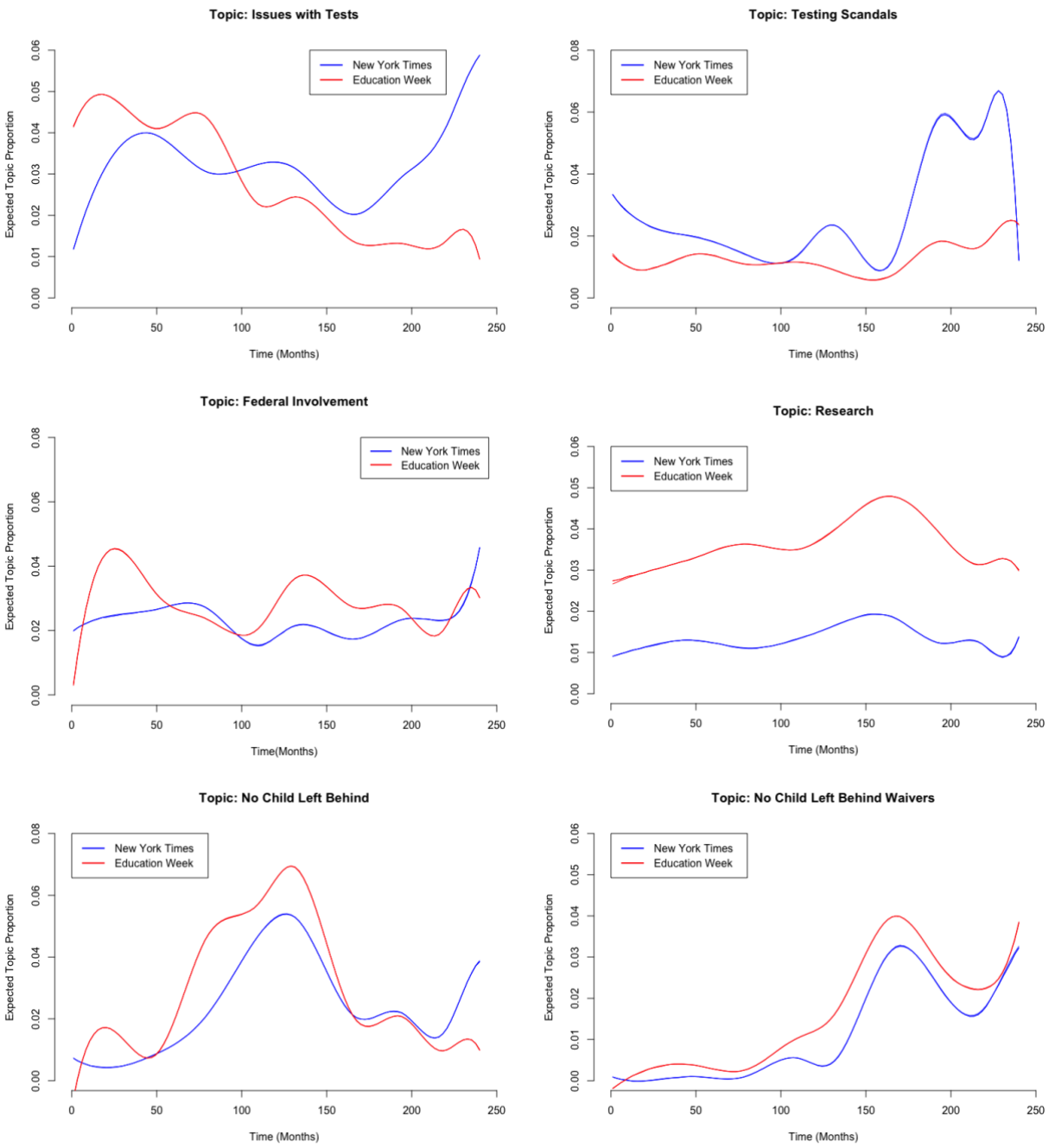
Figure 4.14. Mean Difference in Publication Topic Proportion for Frame Elements



Another method for analyzing differences by publication that provides a more nuanced perspective is to graphically depict proportions of coverage by publication for each frame element individually. Several of the salient frame elements tracked closely together over time across the two publications. For example, the related frame elements on NCLB and NCLB Waivers showed no substantial differences in coverage by publication (see the plots in Figure 4.15).

Figure 4.15. Plots of Proportions by Publication for Positive and Negative Frame Elements





The three frame elements with notable mean differences in Figure 4.14 also showed consistent gaps in the plots of coverage over time. The Research topic had a substantial difference in coverage between the two publications with the proportion of coverage in *Education Week* consistently higher than coverage in the *New York Times* throughout the 20-year period. This finding indicates that the professional publication for educators is more likely to

include articles on education research in its reporting. Because *Education Week* is solely covering education, it may be more likely to cover topics or news in education that would not be considered sufficiently newsworthy to be included in the *New York Times*, which is more likely to be selective in its coverage of education issues, given limited space in the publication. This was further supported by skimming articles highly associated with this topic, which often included details about research and sidebars or hyperlinks to the studies cited. This level of in-depth detail would not be expected in high proportions from a general readership publication such as the *New York Times*.

The higher level of coverage of 4th and 8th grade exams in the *New York Times* can largely be explained by the development of state tests for these grades in New York State in 1996. These tests were developed as a means of determining whether or not schools were successfully educating students and served as a precursor to the testing requirements for NCLB (Medina, 2010). After these state tests were implemented, the *New York Times* regularly reported on the results of the tests, often highlighting student test scores in New York City. An article from September 2005, for example, stated:

> The number of fourth graders performing at grade level in math increased markedly across New York State this year, with the biggest gains in Yonkers and New York City. But the number of eighth graders scoring at grade level declined slightly statewide, ending four years of steady gains in New York City (Herszenhorn, 2005).

The School Performance Ratings topic also exhibited a difference in coverage across time, with the most substantial gap occurring between 1996 and 2003 (see Figure 4.15). This suggests that *Education Week* was covering state-level school performance rating systems in the years before NCLB brought this issue into the national spotlight. For example, *Education Week*

covered California's release of school performance rankings in 2000 but the *New York Times* did not.

For several of the other salient frame elements, there were notable differences in coverage (as seen in the graphical displays in Figure 4.15) even in the absence of an overall mean difference in proportion. For example, coverage of the Common Core State Standards and the surrounding controversy began to increase in 2009 in *Education Week* while coverage in the *New York Times* spiked dramatically in late 2011 into 2012. It may be that *Education Week* was more in tune to early developments around the Common Core and therefore reported on this issue earlier than other media outlets. Several other issues exhibited this general pattern of *Education Week* leading the *New York Times* in terms of coverage. Coverage of both the NCLB and NCLB Waivers topics lagged behind in the *New York Times* during periods when coverage was increasing, suggesting that in some cases, *Education Week* was the leader in reporting on major developments in education (see Figure 4.15).

Another example of differences in coverage between the two publications is the Issues with Tests topic. Coverage of this topic tracked fairly closely during the first two-thirds of the dataset but the two media outlets diverged drastically in coverage during the final years, starting around 2009. Coverage in the *New York Times* increased sharply during this later period, while coverage in *Education Week* continued a trend of slight decline. This difference may be related to coverage of the opt out movement in New York State, where the movement had the largest percentage of families opting out during the last years in the study period in 2014 and 2015.

Similar to the Issues with Testing frame element, the Testing Scandals element also exhibited a notable divergence in the last third of the dataset. *Education Week* had a slight increase in proportional coverage of testing scandals in the final years of the dataset but not to

the extent of the *New York Times*, which exhibited a drastic increase in coverage starting in 2011 (around the time of the Atlanta cheating scandal) that was sustained through the final years of the dataset. This finding may be an indication that general readership newspapers are more likely to cover education, broadly speaking, when there is a controversial or newsworthy issue to report.

Overall, differences in coverage of salient topics between the two publications were minor, which indicates that there were no significant differences in the framing of this issue in the two different media outlets. These two newspapers differed primarily in terms of audience – *Education Week* is a trade publication for educators and the *New York Times* is a general readership publication. Therefore, differences may exist in coverage among media outlets that differ on other dimensions such as political leaning or type of publication (e.g. newspaper or blog).

Overlay of Frames with Testing Chronology

Figure 4.16 provides a chronology of major policy developments and the four frames identified in the current study. In the mid to late 1990s, there were several policy developments at the federal level that contributed to the increasing prominence of testing in schools. The Goals 2000: Educate America Act was signed by President Clinton in spring 1994, offering financial assistance to states that developed plans for standards-based education reform. One aspect of the law was testing in reading and mathematics in designated grades. Later that same year, the Elementary and Secondary Education Act was reauthorized as the Improving America's Schools Act. Together, these two laws placed testing in a central position in educational reform efforts. In addition, these laws laid the foundation for the greater federal role in education that was to come with the passage of NCLB. In 1997, President Clinton proposed a national test for schools.

In his State of the Union speech from that year, Clinton offered the following as part of his call to action for educational reform:

> To help schools meet the standards and measure their progress, we will lead an effort over the next 2 years to develop national tests of student achievement in reading and math. Tonight, I issue a challenge to the Nation: every state should adopt high national standards, and by 1999, every state should test every fourth grader in reading and every eighth grader in math to make sure these standards are met. Raising standards will not be easy, and some of our children will not be able to meet them at first. The point is not to put our children down but to lift them up. Good tests will show us who needs help, what changes in teaching to make, and which schools need to improve (Clinton, 1997).

For Clinton, a national test was part of the solution to the problem of ensuring that all students "succeed in the knowledge economy of the 21st century" (Clinton, 1997). Clinton continued to advocate for the development of a national exam, mentioning it again in his 1999 State of the Union address. Although it ultimately received limited support and did not become federal policy, this proposal further reflected the increasing prominence of testing as a component of school improvement. The Clinton administration also pushed for greater accountability in education later in 1999 as part of proposed legislation to reauthorize the Elementary and Secondary Education Act (Anderson, 1999; Clinton, 1999).

The push for testing on the federal agenda continued with the presidency of George W. Bush. In his first major address before a joint session of Congress in February 2001, Bush described his vision for an increased budget for education and the central role of testing in holding schools accountable:

> Yet when the Federal Government spends tax dollars, we must insist on results. Children should be tested on basic reading and math skills every year between grades three and eight. Measuring is the only way to know whether all our children are learning. And I want to know, because I refuse to leave any child behind in America (Bush, 2001).

The late 1990s and early 2000s also saw an increase in the number of states with exit exam policies (Education Commission of the States, n.d.). These testing policies were often enacted along with new state education standards, which were also gaining in popularity at that time. For example, the development of the California High School Exit Examination (CAHSEE) was aligned to new state-level content standards in English and math (California Department of Education, n.d.). By the end of the 1990s, 27 states had exit exam policies in place. This period of popularity of exit exams overlaps with Frame 1, which is characterized by a positive framing of testing. This primarily positive framing of testing may have reinforced state policymakers' efforts to implement exit exams.

One of the first important developments in the 2000s was the passage of NCLB, which occurred approximately at the end of the first period of framing. During the years from 1996 to 2002, the framing of testing was primarily associated with positive aspects of the issue, including the use of test score data to identify school performance, the use of exit exams as a tool to ensure rigor in schools, and the use of test score data to provide comparable information on the quality of education in states across the U.S. Testing was also proposed as a method to hold schools and teachers accountable to ensure that all students received an adequate education. This more positive framing may have contributed to a policy space that was primed for the annual testing mandate that became a core component of the reauthorization of ESEA as NCLB.

As noted earlier in this chapter, another important period of time in terms of media coverage of testing was the late 2000s and early 2010s, when the gradual shift from Frame 2 to Frame 3 occurred. Several significant policy developments occurred during this time period. First, the Race to the Top grant program was introduced by the White House in the summer of 2009. Race to the Top was a competitive grant program with unprecedented levels of federal funding for states that agreed to implement reforms in four areas: adopt standards for college and career readiness, build data systems to track student achievement, develop methods for identifying effective and ineffective teachers, and institute turnaround models for low-performing schools. U.S. Secretary of Education Arne Duncan, in an editorial in the *Washington Post* announcing the grant program, wrote, "For states, school districts, nonprofits, unions and businesses, Race to the Top is the equivalent of education reform's moon shot" (Duncan, 2009).

The second major policy development that occurred around the time of the transition from Frame 2 to Frame 3 was the introduction of the Common Core State Standards, which were released in the summer of 2010 and subsequently adopted by most states over the following months ("Development Process," n.d.). These new standards required the development of new assessments to measure student performance. Soon after the release of the standards, two state-led consortia began to develop common assessments. In some states, these Common Core-aligned tests have been even more controversial than the standards themselves (Ujifusa, 2014a). School districts, teachers, teachers' unions, parents, and students have all been active in opposition to these new exams (Harris, 2015; Strauss, 2014). In Illinois, for example, a number of school districts joined in opposition to the implementation of PARCC tests (Rado, 2014).

In addition to these policy developments, the Atlanta cheating scandal also occurred during Frame 3. Coupled with the growing backlash to NCLB and the introduction of NCLB

waivers, the policy environment regarding testing leading up to and during Frame 3 was very different from earlier periods. Testing was, by the late 2000s, firmly entrenched as a core component of schooling. And yet, a tension existed during this time between efforts to further solidify testing as a solution to problems such as identifying ineffective teachers (through Race to the Top) and calls to deemphasize testing to reduce the burden and pressure on schools, teachers, and students (NCLB waivers).

During this period, coverage shifted from a higher number of topics associated with positive aspects of testing to less coverage of these topics and a concurrent increase in coverage of topics associated with negative aspects of testing. This trend began with the evolution of the framing of testing from Frame 2 to Frame 3 and continued through the end of the study period.

Although it is not possible to identify the direction or strength of the relationship between media coverage and policy developments, it is interesting to note that during the same period of time when media coverage shifted to less positive aspects of testing, there were also several major shifts in political discourse and policymaking. In 2014 and continuing into 2015, the federal government, which had long been a proponent of testing, released statements proclaiming that testing had become a problem. In October 2015, for example, the U.S. Department of Education released a statement on the administration's plans to reduce over-testing. The statement noted,

> In too many schools, there is unnecessary testing and not enough clarity of purpose applied to the task of assessing students, consuming too much instructional time and creating undue stress for educators and students. The Administration bears some of the responsibility for this, and we are committed to being part of the solution (U.S. Department of Education, 2015a).

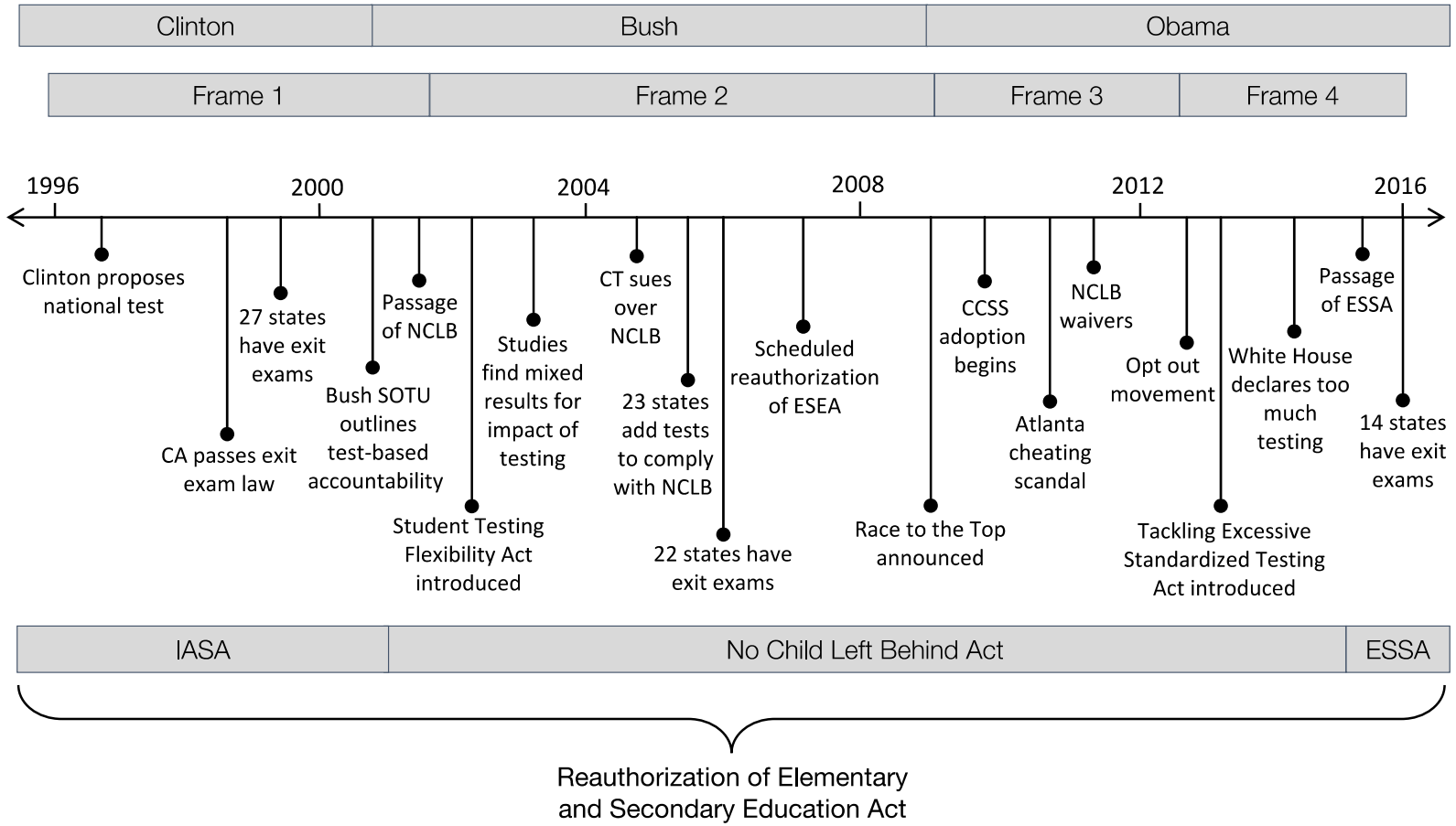
The statement outlined specific actions the administration would take to address the ways in which the Administration may have contributed to the problem of overemphasis on testing. These actions included providing financial support for states to develop less burdensome tests, providing guidance and technical assistance to states and districts to reduce testing time, grant flexibility from federal mandates and support for innovative testing practices, and reducing the reliance on test scores through executive actions (U.S. Department of Education, 2015a). The statement also called on Congress to reduce the role of testing in the reauthorization of ESEA. Given the federal government's history of support for test-based accountability and its use of testing as a policy lever for educational reform, this shift in position was particularly notable.

The ongoing coverage of the high-profile cheating scandal in the Atlanta Public Schools system and the growing opt out movement added to the flurry of activity related to testing during this period of time. Although the Atlanta cheating scandal first gained significant media coverage in 2010, newspapers continued to cover the story while trials and convictions played out in subsequent years. This continual coverage is perhaps not surprising, given that the story contained some sensational details. For example, the scope of the cheating scandal was unprecedented. Investigators reported that cheating had occurred across 44 schools and 178 teachers and principals were involved (Office of the Governor, State of Georgia, 2011). Media coverage highlighted particularly shocking descriptions of details of the scandal such as a teacher being forced by a principal to crawl under a desk because of low test scores or a teacher noting that the district was "run like the mob" (Turner, 2011). This media coverage kept the controversy about testing in the spotlight for several years. For example, a 2013 article from the *New York Times* was titled "Atlanta Cheating Scandal Reignites Testing Debate." Similarly, in 2014, *Education Week* published an article titled "Atlanta Cheating Scandal Lessons" in which

the author noted that when test scores are used for "naming and shaming" teachers, more of these kinds of scandals can be expected (Gardner, 2014).

The impact of these developments is difficult to determine but it did not result in the elimination of testing as a central aspect of the reauthorization of the Elementary and Secondary Education Act in 2015. Although states were given more flexibility under the new law, the annual testing requirement that was a core component of NCLB remained. However, the increased flexibility given to states under ESSA effectively decoupled testing and high-stakes accountability. The new law eliminates the mandate that schools must achieve "adequate yearly progress" and allows states to establish their own accountability goals for addressing performance issues. Additionally, state-established accountability systems must include at least one non-academic indicator that is not related to test scores, such as school climate or student engagement (Klein, 2016; OBrien, 2016).

Figure 4.16. Chronology of Testing Developments, 1996 – 2016



Note. CCSS = Common Core State Standards, ESSA = Every Student Succeeds Act, IASA = Improving America's Schools Act, NCLB = No Child Left Behind Act

Conclusion

The analysis of frames and frame elements from the topic modeling suggests that early frames in the media coverage of testing focused on more of the positive aspects of testing, including measuring student learning (4th/8th grade tests, trends in test scores), comparable data on performance (NAEP), providing families and public with information (school performance ratings), and establishing high expectations for students and educators (Regents exams, high school exit exams). Later frames (covering the period from 2009 through 2015) included substantially fewer of these positive elements of testing and included more controversial and negative frame elements (such as testing scandals and the controversy surrounding the Common Core and testing), which shifted these later frames to more critical coverage of testing. The analysis of the framing of the issue suggests two trends. The first trend was a gradual decline in coverage of topics associated with positive aspects of testing. The second trend was a gradual increase in coverage of topics associated with negative aspects of testing. Taken together, these trends in the data suggest a slow evolution over time as many of the positive frame elements were dropped and new, more negative frame elements became salient. This evolution led to a reframing of the issue by the last period of time that was much more negative than the initial framing.

Overall, the analysis suggests that coverage of testing in the past 20 years has been closely tied to many other important, and often controversial, aspects of education policy, including school performance ratings, national standards, and teacher merit pay. Additionally, the last frame, covering a period from 2013 to 2016, focused on more controversial aspects of testing than previous frames with the increased coverage of the Common Core State Standards and testing scandals. The evidence suggests a shift from a mostly positive framing of testing to a

primarily negative framing of testing over the time period studied. In Chapter Five, I revisit the research questions and discuss the findings, significance, and implications of this study.

CHAPTER 5: DISCUSSION

This chapter summarizes and discusses the findings from the study; discusses the significance and implications of the research, including an assessment of the conceptual framework and the application of topic modeling; and discusses limitations and future directions for research.

Summary Review

In this study, I set out to examine media coverage of testing in schools in two newspapers across a 20-year period. I sought to understand how the media frame the issue of testing by selecting and highlighting certain dimensions of the issue and how these salient dimensions change over time. I searched the archives of *Education Week* and the *New York Times* for all articles that included discussions of testing in schools from 1996 through 2015, developing a dataset of over 8,400 articles. Using a text analytics technique called structural topic modeling, I analyzed this collection of articles to identify topics in coverage of the issue of testing. The final model used in the study estimated 38 topics. I then analyzed patterns of relatively higher salience topics to identify frame elements. Clusters of these frame elements indicated a frame. A total of four frames were identified across the 20-year period. An examination of these frames indicated a gradual but consistent shift from more coverage of positive issues related to testing to more coverage of negative issues related to testing. Two trends contributed to this shift. The first trend was that earlier frames had a higher number of frame elements associated with positive aspects of testing and this number decreased over time until the final frame had

none of these positive frame elements. The second trend was that later frames had a higher number of frame elements associated with negative aspects of testing.

Research Question 1

The first research question asked how the issue of testing in schools has been framed in media coverage. Using structural topic modeling, a total of 38 topics related to testing were identified in the dataset. The topic model suggests that testing is related to a variety of other issues in education, covering everything from math and reading to gender issues and English Language Learners. This finding of a high number of topics related to testing was not surprising, given the complexity of the underlying issue. The next step was to identify topics with relatively high salience at any point in time in the dataset. By examining the distribution of all topic proportions across time, I was able to establish a level for high salience relative to the entire topic model. Through this process, I identified a subset of 19 topics exhibiting this high salience. By identifying topics with higher salience in the corpus and labeling these as frame elements, I was able to gain an understanding of how media coverage shaped the issue space. I identified four frames in the 20-year period of the study. Frame 1 covered the period from 1996 to 2002, Frame 2 covered the period from 2002 to 2009, Frame 3 lasted from 2009 to 2012, and Frame 4 lasted from 2013 to the end of the study period at the end of 2015. There were substantial changes in the composition of the frames (as newly salient frame elements were added and other frame elements dropped in salience) to identify these periods of time as distinct frames, but the process of frame change is best described as gradual and evolutionary. That is, in the three transitions to new frames, there were at least three frame elements in each instance that carried over from the previous frame. Salient frame elements were then categorized according to whether they were positively, neutrally, or negatively associated with testing.

Research Questions 2 and 3

The second and third research questions asked how the debate over testing in schools has evolved over time and to what extent certain dimensions of the issue dominate coverage at different periods of time. Findings from the study indicate that there was an evolution of the issue over time and that salient elements also varied over time. Specifically, the framing of the issue evolved from primarily positive in the first 13 years to primarily negative by the last 3 years of the study. In keeping with previous research on issue framing, however, the findings from the study also suggest that frames are quite stable. The shift from a more positive framing of testing to a more negative framing occurred gradually over the twenty-year period and, as noted above, some frame elements were consistent through several shifts in framing. Additionally, the first two frames identified in the study lasted for approximately six and seven years. Overall, therefore, media coverage remained relatively stable but with a sufficient number of shifts over time to ultimately lead to a more substantial evolution by the end of the study period.

Research Question 4

The fourth research question sought to determine whether or not the frames in media coverage differ in general versus professional newspapers. In order to answer this question, I included articles from two publications - one professional newspaper for educators (*Education Week*) and one general readership newspaper (*New York Times*). One of the features of structural topic modeling is the ability to include covariates in the model. I was therefore able to include a publication covariate to distinguish articles published in *Education Week* and articles published in the *New York Times* in the topic model. This feature allowed me to examine mean differences in topic proportions by publication and also to plot differences in proportions by publication type

across time. The analysis by publication type revealed that, overall, there were only minor differences in coverage of salient topics between the two publications, which indicates that there were no significant differences in the framing of this issue in the two different newspapers.

Discussion of Findings

In addition to providing insights to the research questions outlined above, this study had several notable findings that are discussed below. These include the nature of the issue space in the years leading to passage of NCLB, the apparent shift triggered by NCLB, the high salience of the one explicitly negative topic related to testing, and the absence of an explicitly positive topic.

The first frame identified in the study covered the period of years from 1996 leading up to the passage of NCLB in 2002. This frame contained the highest proportion of frame elements associated with positive aspects of testing. For example, frame elements such as coverage of trends in test scores and school performance ratings highlight the importance of tests as a tool to gain information about students, schools, and the state of education in the U.S. The composition of this first frame suggests that the issue space was conducive to the passage of NCLB with its focus on annual testing. When President George W. Bush signed the law in January 2002, he stated:

> The first way to solve a problem is to diagnose it. And so, what this bill says, it says every child can learn. And we want to know early, before it's too late, whether or not a child has a problem in learning. I understand taking tests aren't fun. Too bad. We need to know in America. We need to know whether or not children have got the basic education (Bush, 2002).

This "need to know" how students were performing had already been established in the six years prior to the passage of the law through the framing of testing in media coverage.

Although the issue space in the years leading up to 2002 may have influenced the passage of NCLB by being primarily positive toward testing, the actual passage and implementation of the law appears to have triggered a slow but steady shift to more negative coverage. As depicted in Figure 4.13, coverage of NCLB reached a peak around 2007 and in subsequent years several topics that were associated with negative aspects of testing increased in coverage. This finding suggests that the implementation of NCLB and debates about the impact of the law may have induced the shift in framing in later years of the study period. In the years after passage of NCLB, other developments further reinforced this shift to a more negative media portrayal of testing in schools. The high-profile Atlanta testing scandal, for example, ignited a national debate about the extent to which teachers and administrators or the high-stakes tests were to blame for the widespread cheating that occurred. Coupled with the growing dissatisfaction with NCLB, increasingly negative coverage of testing may have contributed to the shift in position on testing by the federal government in the last years of the study period.

As noted in Chapter Four, one explicitly negative topic was present in the topic model. The Issues with Tests topic was a salient frame element in three of the four frames. However, there was an overall drop in proportional coverage of this topic over the 20-year period of the study (the topic proportion reached a high in 1999 and then steadily dropped to a low around 2010 before gradually beginning to increase again). There are several possible explanations for why this decrease over time occurred. One explanation is that in the early years of the study, the Federal Involvement topic was also highly salient and this topic included coverage of the Clinton administration's push to implement a national test. Pushing for a national test on the federal agenda may have prompted increased discussions about issues with tests. Another possible explanation is that during this early period of years, NAEP coverage was

also particularly high. Comparisons are often made between NAEP scores and state assessment scores and these comparisons may have led to increased discussions about the tests, including questions of proficiency levels, test design, and so on. Although a detailed analysis of the relationships between topics is beyond the scope of this study, future research might help to unpack some of these trends in media coverage by examining in greater detail how these issues are related.

Another interesting aspect of the findings was the absence of an explicitly positive topic. As discussed previously, the Issues with Tests topic was the only explicitly negative topic in the dataset. However, there was no complementary but contrasting topic explicitly about the positive aspects of testing. It may be that testing has been a core part of schooling for so many years that positive aspects of testing are seen as a given or are not considered interesting enough to warrant any substantial media coverage. Another possible explanation is that scholars such as Richard Phelps and other proponents of testing are correct in asserting a media bias in favor of critics of testing. This explanation is supported by the finding that the Issues with Tests topic had the highest overall proportion of coverage of any topic in the study (see Figure 4.2). However, the framing of the issue described in this study suggests that when the frame elements are examined as a whole, there was no clear indication of overall bias in either direction, although coverage did shift from a more positive to a more negative framing. A third possible explanation is that as accountability pressures increased over the past 20 years, tests have increasingly been put in the spotlight and scrutinized. For example, if test score data are being used to evaluate teachers and inform decisions about employment, then potential problems with tests become a more important issue. Finally, it may be that proponents of testing have not

been as vocal in advocating for the benefits of testing as critics have been in pointing out potential flaws with testing.

The findings from the 20-year period of the study suggest that the current framing of testing is likely to continue. One reason this is likely is that the final frame in the dataset (Frame 4) contained no positive frame elements. Given that the shift to a negative framing of testing occurred gradually over the 20-yeard period, a rapid shift back to a more positive framing of testing is unlikely. Additionally, media coverage of recent policy action at both the federal and state levels to limit testing combined with the growing opt out movement may be contributing to a self-reinforcing process in which the shift in understanding of the role of testing in schools (and shift in policy position) further reinforces the negative framing of the issue.

In Chapter Two, I discussed the use of tests as policy instruments. McDonnell (2005) noted that tests were an appealing and popular instrument for politicians as a means of influencing classroom practice. For politicians, testing served as a policy solution to problems in schools such as social promotion, inequity among subgroups of students, and underperforming teachers. Advocating for testing was a way for politicians to communicate a hard-line stance on education through accountability. The rhetoric of George W. Bush on the unacceptability of leaving any child behind and the necessity of testing to ensure this didn't occur is one example of this tough-on-schools stance. However, the shift in the framing of testing found in this study suggests that tests are less likely to be an appealing policy instrument to politicians now that negative aspects of the issue are more salient than positive aspects.

Significance

This study is a novel and innovative contribution to the literature in several ways. First, there are few applications of topic modeling to education policy research. The current study

provides one example of how this method can be productively applied to answer questions in education policy that would be very resource intensive to examine using traditional coding techniques. Specifically, the application of structural topic modeling to study issue framing in media coverage of an education policy issue was productive and resulted in a greater understanding of how the testing issue has been framed and how this framing has changed over time.

Second, the conceptual framework employed in this study builds on previous work on the theory of issue framing. In particular, empirical scholarship on measuring issue dimensions and frames such as the work of Baumgartner, DeBoef, and Boydstun (2008); Mathres and Kohring (2008); and Nowlin (2016) informed the development of the conceptual framework outlined in Chapter Two. As scholars have long noted, conducting empirical work on issue framing is challenging. It is therefore important to continue to develop and refine techniques for identifying and measuring frames and to apply and assess these techniques in various disciplines. The current study sought to contribute to this important work by developing a conceptual framework and combining it with an innovative big data analytic approach to study issue framing in education.

Third, media coverage influences the perspectives of the public and policymakers. As DiMaggio, Nag, and Blei (2013) noted, "Press coverage both reflects and represents one stream of influence in the formation of elite and public opinion" (p. 574). In order to gain a better and more comprehensive understanding of the forces that shape public and elite perceptions and ultimately lead to policy change, it is therefore important to understand the role that media play in framing issues in education policy. A lack of research on the role that media play in education generally is problematic. Without a better understanding of how media portray issues in

education and how these portrayals can shift over time, an influential factor in policymaking and policy change remains unexamined. The education research community, in particular, would benefit from a better understanding of the ways in which media frame issues, interpret research, and engage in education policy debates. Engaging in public scholarship and facilitating the use of research in policy and practice is a critical aspect of education research. As such, a greater understanding of the role of media in education policy can inform efforts to engage in this public discourse. This study also fills a gap in the literature on the politics of testing. There are no previous empirical studies of media framing in the testing debate. The current study therefore opens up a space for thinking about the issue of testing and media coverage.

Given previous scholarship indicating a lack of media coverage of education generally (see Chapter Two), the relatively high number of articles in the *New York Times* that included coverage of testing in education was unexpected. The dataset included almost 3,000 articles from the *New York Times* during the 20-year period, which is an average of approximately 3 articles each week related to this issue. One possible explanation for this high amount of coverage in a general readership newspaper is that testing is related to numerous other issues in education policy and therefore is likely to be included in coverage ranging from the effectiveness of charter schools to teacher merit pay to college readiness. This is supported by the finding of a large number of topics in the model output, which included a wide range of related topics. It may be that as education becomes an increasingly partisan and controversial issue in U.S. politics and policy, media coverage of education policy generally and testing in particular will increase. This further supports the need to better understand how the media frame issues in education policy, as growth in media coverage may result in a greater influence of the media on education policy.

Implications

Assessing the Conceptual Model

In the current study, analyses of changes in frame elements (salient topics) was helpful in identifying higher level frames and revealing how this framing changed over time. That is, the conceptual framework outlined in Chapter Two was appropriate for identifying topics as frame elements and tracking changes in frames over time. The conceptual framework also was appropriate for identify the resonance and persistence of frames. This aspect of the model was beneficial in the analysis as it indicated that Frame 1, which was the most positive toward testing of the four frames, was particularly strong and therefore may have associated with the passage of NCLB.

However, it is important to note that complex issues such as testing may benefit from an analysis at a more granular level of analysis than is possible using an unsupervised method such as topic modeling. Although the current study is an important step in applying innovative methods for identifying frame elements and frames in education policy, additional developments in topic modeling might allow for more nuanced identification techniques that are still able to process a large corpus of documents but that account more directly for arguments in the debate. The frame elements identified in the current study can best be described as associations between the focal issue of testing and related topics rather than explicit arguments for or against testing.

The current study provides insights into the contexts (topics) in which testing is discussed and, from the interpretation of these contexts, it is then possible to explore how arguments associated with certain topics are more or less salient over time. In the conceptual model, these clusters of topics are interpreted as frame elements that lead to an overarching framing of the

issue. The analysis suggests that the framing of the issue does shift over time. In that sense, the conceptual framework was productively applied to gain a better understanding of the framing of testing in media coverage by identifying a gradual but consistent shift in framing over the 20-year period of the study. However, the current study did not identify specific arguments in the testing debate, which is partly due to the nature of topic modeling. The use of topic modeling for this type of study is discussed in the next section.

The conceptual model with distinct frames clearly demarcated across time serves an important analytical purpose but the findings suggest that the framing of the issue in media coverage is a much more fluid process of change, as illustrated by the continuity of multiple frame elements across time (and therefore across frames). This finding of a fluid process of change is likely a more accurate representation of how media coverage changes over time (with the exception of major individual events that result in a spike of coverage on a particular topic for a relatively short period of time), which occurs slowly as the dimensions of the issue evolve over time.

Assessing the Application of Topic Modeling

As with any method for conducting research, there were both advantages and disadvantages to selecting topic modeling for the current study. The application of topic modeling for an analysis of issue framing in media coverage has three advantages. First, previous scholarship on issue framing has noted the general stability of the frames associated with public policies and the difficulty of reframing issues (Baumgartner, DeBoef, & Boydstun, 2008). It is therefore important to utilize techniques that can account for relatively long periods of time. Topic modeling provides an opportunity to leverage a big data text analytic technique to examine a large corpus of documents across many years. In order to adequately assess the extent

to which the framing of the issue changed over time, it was necessary to examine a lengthy period of time. The use of topic modeling described in this study allowed for a breadth of analysis that would otherwise be limited by resource and time constraints. This feature of the approach was particularly important in retrospect, as the findings from the analysis suggest that the framing of testing shifted gradually over the 20-year period. A study covering a more limited period of time might miss these shifts in framing.

Second, topic modeling provides the ability to identify multiple topics within a document. For multidimensional issues such as testing, this feature of the method is a distinct advantage. Rather than coding articles as a whole, topic modeling identifies the topics embedded in each article, thereby providing an additional level of detail. By coding articles at this level of detail, topic modeling has an advantage over text analysis techniques that code at the level of the entire article. This means that topic coverage is likely to be more comprehensive, as even topics that are embedded within articles that are primarily about another topic are identified.

Third, the sheer size of the dataset is another advantage. In the absence of automated text analysis techniques, the analysis of media coverage would likely require building a dataset that consisted of a random sample of articles from each year included in the study period. Applying a big data technique provided the opportunity to include all relevant articles from the study period, which in turn led to the ability to identify more nuanced changes in the dataset than might be present in a random sample of articles across years. By analyzing all articles in the two media outlets over the twenty-year period (over 8,000 articles), concerns related to sampling are reduced.

There are also several limitations to topic modeling. One is that the topics identified in the model were broader than arguments. It was therefore necessary to make inferences about

whether these related topics were positively associated with testing, neutral to testing, or negatively associated with testing. If identifying the precise arguments in the debate is the primary goal of the research, other methods may be better suited than topic modeling to achieve this goal. Another limitation of topic modeling is that as an unsupervised learning method, all interpretation occurs post-estimation. That is, the processes of interpreting topics and determining whether they are positive, neutral, or negative happens after modeling. Topic modeling does not allow for assigning these values to the data prior to estimation. Although this can be a limitation if analysts have developed a priori frameworks for the data, unsupervised models have benefits as well and one of the primary ones is the ability of the models to discover topics that might not otherwise be analyzed.

This study highlights the potential use of topic modeling as a strategy for identifying and analyzing topics in a large corpus of documents within the field of education on any number of issues, including, for example, teacher hiring practices (using job application data), education finance (budget documents), and public discourse on the Common Core State Standards (social media posts, media coverage), and many other areas.

Limitations

This study was limited in the following ways. First, identifying the appropriate number of topics to include in the final model is a decision made by the analyst that can potentially alter the interpretation and findings. Using the same dataset, an analyst that determined that 20 topics was the appropriate parameter for the model would have a different study than an analyst that determined that 50 topics was the appropriate number. Second, the current study was limited to press coverage in two newspapers. It is therefore difficult to generalize the findings from this study to public coverage of this issue more generally, including other newspapers or other types

of media such as blogs or Facebook posts. Public and policymaker perspectives are shaped by coverage of events from numerous sources well beyond newspaper coverage. A more complete picture of the public discourse related to the issue of testing would need to include additional sources, both newspaper coverage and new media. A related limitation is the extent to which coverage may differ by newspapers depending on political leaning. The *New York Times* is often characterized as a liberal leaning newspaper and it is therefore possible that coverage would differ systematically in some important way from coverage in other newspapers. As noted in Chapter Three, research suggests that the media agenda remains stable regardless of political leaning of particular newspapers, but this finding has not been tested on coverage of education policy issues. Future research might include additional general readership newspapers to assess the extent to which political leaning impacts coverage and framing.

Future Research

Future research in several directions would provide additional insights related to the findings in this study. One area where this research could be expanded is in the number and types of media outlets included in an analysis. New media such as Twitter and Facebook are increasingly influential in both public and elite discourse and perceptions. A future study might incorporate data from these sources to provide a more comprehensive picture of the influential forces in media and how the framing of issues may differ across types of media. Future research might also include a larger number of newspapers with different political leanings to assess the extent to which framing of issues in education is influenced by these differences.

In addition to different media sources, another direction for future research is to expand the type of documents analyzed beyond media. Media coverage is only one part of a rich record of data on a policy issue. Documents such as Congressional hearings, presidential speeches, or

bill texts could also be analyzed to provide a broader perspective on the framing of the issue by various actors in the policy debate.

As discussed earlier in this chapter, findings from this study suggest that the passage of NCLB may have prompted the shift to more negative coverage of testing. Given the importance of NCLB generally as a policy development in education and in relation to the issue of testing, another direction for future research is an in-depth analysis of the framing of the law in media coverage. For example, researchers might examine the extent to which coverage of the law itself shifted to a more negative tone. Although the current study examined NCLB only within the context of testing, the findings suggest that NCLB coverage did shift, particularly once waivers were introduced. Additionally, a study designed to analyze the framing of the law in media coverage might help to explain why Congress struggled to reauthorize the law for many years.

There are also ways in which future research would add to the literature on frame analysis. Future applications of topic modeling to education policy issues will help to provide additional clarification about the kinds of research questions that are conducive to topic modeling and questions that more traditional methods of analysis are better suited to answer. Additionally, applying the conceptual framework in this study, analysts could further test the use of topic modeling to answer questions about frames and frame elements.

Finally, quantitative text analysis is a developing field. As such, new and innovative developments will help to further refine and build on techniques such as topic modeling. As more work is done in this field of research, the methods will continue to improve and the potential applications will continue to expand.

Testing in Schools

Testing in schools has a long and controversial history. In the last two decades, however, as tests were increasingly tied to accountability and the stakes of these tests increased, public discourse on this issue has changed. Media coverage, specifically, has evolved over the two-decade period of this study. As the testing solution gradually evolved into the testing problem, political and public activity also was evolving. The opt-out movement, state- and federal-level policy action, and, ultimately, the passage in 2015 of the Every Student Succeeds Act all signal a shift in the understanding of the appropriate place of tests in schools within the past decade.

The debate is unlikely to dissipate, even with state and federal action to limit testing in schools. As discussed in Chapter Two, testing has always been a perennial issue in education and even with current attempts to reduce the amount of testing and design better tests, these efforts are unlikely to put an end to the debate about the proper role of tests in schools.

This debate reflects the larger tension about education that exists today. The increased national attention to education reflects a larger reality of American society: the public school is one of the few remaining social institutions where all segments of the population come together. What we choose to do in these schools, therefore, reflects our values and goals for society. Our society has placed immense significance on schooling, leading to an overarching narrative that education is the key to individual health and collective economic well-being. For the past twenty years, testing and test-based accountability have played a central role in this reality, which lead to questions such as: What role, if any, should tests play in education accountability systems? Do tests help us to create a more egalitarian educational system that ensures that no child is left behind? Do they foster academic rigor and adherence to high standards for what all students should know and be able to do? These are complex, difficult questions. As we put more

pressure on our system of education as the means to ensure the health and prosperity of our nation, controversial aspects of schooling including testing will continue to be debated. Therefore, a greater understanding of how the nation thinks about testing can lead to a greater understanding of how the nation thinks about education generally. Scholars of education policy must continue to investigate the role of educational tests and test-based accountability, how testing in schools is framed in public discourse, and other aspects of the testing debate to contribute to a better understanding of education in our society and an improved way forward for students and schools.

APPENDIX: LIST OF TOPICS AND ASSOCIATED TERMS

Topic Number	Topic	Terms	Frequent and Exclusive Terms
1	Technology	compani, technolog, onlin, educ, comput, servic, use	softwar, erat, edtech, app, sylvan, digit, compani
2	Math	student, math, class, high, cours, art, take	algebra, art, saturday, geometri, junior, cours, dualenrol
3	Superintendent/Principal Accountability	district, princip, superintend, counti, year, administr, leader	princip, counti, district, superintend, vernon, leadership, staff
4	School Performance Ratings	school, improv, account, system, achiev, perform, student	account, improv, card, target, school, lowperform, perform
5	Regents Exams	state, will, board, new, regent, offici, commission	regent, edison, commission, rhode, portfolio, island, statewid
6	School Law and Disability	disabl, court, special, rule, educ, student, case	court, plaintiff, suprem, lawsuit, justic, disabl, iep
7	No Child Left Behind (NCLB)	law, child, left, behind, feder, educ, act	nclb, ayp, law, spell, left, behind, paig
8	Public Opinion	percent, year, texa, survey, florida, public, half	survey, poll, fcat, romney, texa, percent, homeschool
11	Achievement Gaps	gap, black, white, minor, hispan, student, achiev	hispan, black, africanamerican, racial, latino, gap, white
12	Teacher Merit Pay	teacher, union, evalu, teach, pay, system, year	bonus, salari, certif, union, licens, compens, teacher
14	High School Exit Exams	student, exam, pass, graduat, high, requir, state	ged, diploma, exit, pass, exam, graduat, mill

15	Race to the Top	educ, grant, top, reform, race, feder, million	stimulus, economicstimulus, grant, race, winner, reform, top
16	Literacy Instruction/Coaching	teacher, teach, instruct, classroom, learn, read, lesson	lesson, literaci, classroom, workshop, instruct, teach, coach
18	Gender Issues	univers, girl, program, job, candid, train, engin	women, girl, teachercandid, male, femal, gender, tfa
19	International Comparisons	scienc, countri, nation, american, intern, unit, educ	pisa, timss, finland, korea, singapor, oecd, china
20	Grade Promotion Tests	children, read, program, summer, grade, year, parent	kindergarten, summer, kindergartn, children, memphi, danc, harlem
22	State-Level Elections	governor, state, gov, educ, polit, legisl, support	ballot, gov, nea, governor, elect, voter, oppon
23	NCLB Waivers	state, duncan, waiver, depart, evalu, new, system	kline, waiver, duncan, arn, jindal, louisiana, tennesse
24	4th/8th Grade Exams	grade, student, grader, level, score, state, perform	grader, eighth, fourth, grade, fourthgrad, level, eighthgrad
26	State Data Systems	report, data, system, rate, inform, use, educ	data, report, inform, collect, lms, databas, rate
27	National Assessment of Educational Progress (NAEP)	nation, naep, assess, read, profici, progress, math	naep, nagb, sampl, nces, nation, voluntari, profici
29	Federal/State Conflict	state, feder, depart, educ, requir, offici, year	connecticut, regul, compli, subgroup, titl, utah, depart
32	Social Promotion	presid, nation, polici, group, plan, promot, propos	aft, crew, retent, weingarten, promot, yesterday, forum
33	Research	studi, research, found, univers, effect, find, professor	valuead, research, studi, random, stanford, found, conclud

34	English Language Learners (ELL)	english, languag, student, englishlanguag, learner, educ, california	ell, esl, lep, bilingu, englishlearn, englishlanguageprofici, learner
35	Testing Scandals	cheat, investig, offici, contract, answer, educ, problem	cheat, alleg, scandal, tamper, erasur, inspector, investig
36	NYC-specific	citi, new, york, mayor, chancellor, system, bloomberg	mayor, bloomberg, klein, giuliani, chancellor, levi, millionstud
37	Charter Schools	school, charter, public, privat, choic, educ, oper	charter, kipp, rhee, orlean, cathol, parenttrigg, sector
38	Issues with Tests	test, score, assess, student, exam, use, measur	test, error, fairtest, highstak, administ, valid, ctbmcgrawhil
39	College Readiness	colleg, high, student, cours, higher, graduat, educ	postsecondari, collegereadi, vocat, placement, collegelevel, colleg, workplac
41	Federal Involvement	educ, bill, hous, bush, senat, republican, presid	sen, senat, mccain, rep, bipartisan, goodl, bill
42	Common Core Tests	board, assess, will, member, massachusett, accommod, balanc	parcc, mcas, consortia, accommod, smarter, consortium, balanc
43	Education Finance	money, fund, budget, program, spend, million, year	voucher, tax, budget, fiscal, tuition, spend, cut
44	Writing Tests	student, question, write, skill, use, assess, answer	format, vocabulari, reader, passag, write, essay, cognit
45	College Entrance Exams	colleg, sat, act, admiss, score, student, board	collegeadmiss, collegeentr, admiss, sat, verbal, caperton, psat
46	Trends in Test Scores	score, point, year, read, averag, math, gain	percentag, gain, averag, rose, slight, percentil, proport
47	Standards	standard, state, new, curriculum, math, adopt, set	standard, materi, kentucki, commonstandard, nctm, rigor, curriculum

48	Common Core Controversy	common, core, adopt, commoncor, will, art, implement	commoncor, core, common, shawne, draft, align, coleman

REFERENCES

Abedi, J. (2005). Issues and consequences for English language learners. In J. L. Herman & E. Haertel (Eds.), *Uses and misuses of data in accountability testing* (pp. 175-198). Malden, MA: Blackwell.

Adams, C. F. (1880, November). Scientific common-school education. *Harper's New Monthly Magazine*, 61(366), 934-942.

AERA. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452.

Albaugh, Q., Soroka, S., Joly, J., Loewen, P., Sevenans, J., & Walgrave, S. (2014, June). *Comparing and combining machine learning and dictionary-based approaches to topic coding*. Paper presented at the Comparative Agendas Project conference, Antwerp, Belgium.

AlSumait, L., Barbara, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of LDA generative models. Paper presented at the European Conference on Machine Learning, Bled, Slovenia.

Amrein, A. L., & Berliner, D. C. (2002) High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1-74.

Anderson, G. L. (2007). Media's impact on educational policies and practices: Political spectacle and social control. *Peabody Journal of Education*, 82(1), 103-120.

Anderson, N. (1999, May 20). Clinton details education reform plan. *Los Angeles Times*. Retrieved from http://articles.latimes.com

Apperson, J., Bueno, C., Sass, T. (2016). Do the cheated ever prosper? The long-run effects of test-score manipulation by teachers on student outcomes. (CALDER Working Paper No. 155). Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research.

Applebome, P. (1997, February 28). National tests show students have improved in math. *New York Times*. Retrieved from http://www.nytimes.com

Atkinson, M. L., Lovett, J., & Baumgartner, F. R. (2014). Measuring the media agenda. *Political Communication, 31*(2), 355-380.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *The policy effects of PISA*. Oxford: Oxford University Centre for Educational Assessment.

Baumgartner, F. R., De Boef, S. L., & Boydstun, A. E. (2008). *The decline of the death penalty and the discovery of innocence*. Cambridge: Cambridge University Press.

Baumgartner, F. R. & Jones, B. D. (2005). *The politics of attention: How government prioritizes problems*. Chicago, IL: University of Chicago Press.

Baumgartner, F. R. & Jones, B. D. (2009). *Agendas and instability in American politics* (2nd ed.). Chicago, IL: University of Chicago Press.

Baumgartner, F. R., & Mahoney, C. (2008) The two faces of framing: Individual-level framing and collective issue definition in the European Union. *European Union Politics, 9*(3), 435-449.

Behuniak, P. (2003). Educational assessment in an era of accountability. In J. Wall & G. Walz (Eds.), Measuring up: Assessment issues for teachers, counselors, and administrators (pp. 335-347). Greensboro, NC: Caps Press.

Behr, R. L., & Iyengar, S. (1985). Television news, real-world cues, and changes in the public agenda. *Public Opinion Quarterly, 49*(1), 38-57.

Belfield, C., & Crosta, P. A. (2012). Predicting success in college: The importance of placement tests and high school transcripts. (CCRC Working Paper No. 42). New York, NY: Columbia University, Teachers College, Community College Research Center.

Berliner, D. C., & Biddle, B. J. (1999). The awful alliance of the media and public school critics. *The Education Digest, 64*(5), 4–10.

"Big Data Meets Big Data Analytics." (n.d.) SAS White Paper.

Blad, E. (2016, January 6). ESSA law broadens definition of school success. *Education Week.* http://www.edweek.org.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Blei, D. & Lafferty, J. (2009). Topic Models. In A. Srivastava and M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal, 42*(2), 231-268.

Bowers, A. J., & Chen, J. (2015). Ask and ye shall receive? Automated text mining of Michigan capital facility finance bond election proposals to identify which topics are associated with bond passage and voter turnout. *Journal of Education Finance, 41*(2), 164-196.

Boydstun, A. E. (2013). *Making the news: Politics, the media, and agenda setting*. Chicago, IL: University of Chicago Press.

Boydstun, A. E., & Glazier, R. A. (2013). A two-tiered method for identifying trends in media framing of policy issues: The case of the war on terror. *The Policy Studies Journal, 41*(4), 706-735.

Bracey, G. W. (1995). *Final exam: A study of the perpetual scrutiny of American education*. Bloomington, IN: Technos Press.

Brett, M. R. (2012). Topic modeling: A basic introduction. *Journal of Digital Humanities, 2*(1).

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. *Oxford Review of Education, 39*(1), 52-71.

Brown, E. (2015, October 28). U.S. student performance slips on national test. *The Washington Post*. Retrieved from https://www.washingtonpost.com/

Brown, A. B., & Clift, J. W. (2010). The unequal effect of adequate yearly progress: Evidence from school visits. *American Educational Research Journal, 47*(4), 774-798.

Burnette, D. (2016, April 12). Leaders in N.Y. flip-flop on Common Core, opt-outs. *Education Week*. http://www.edweek.org.

Bush, G. W. (2001). Address before a joint session of the Congress on administration goals. The American Presidency Project. Retrieved from http://www.presidency.ucsb.edu

Bush, G. W. (2002). Remarks on signing the No Child Left Behind Act of 2001 in Hamilton, Ohio. The American Presidency Project. Retrieved from http://www.presidency.ucsb.edu

Cacciatore, M. A., Scheufele, D. A, & Iyengar, S. (2016). The end of framing as we know it … and the future of media effects. *Mass Communication and Theory, 19*(1), 7-23.

California Department of Education. (n.d.). California High School Exit Examination (CAHSEE). Retrieved from http://www.cde.ca.gov/ta/tg/hs/

Camara, W. J., & Shaw, E. J. (2012). The media and educational testing: In pursuit of the truth or in pursuit of a good story? *Educational Measurement: Issues and Practice, 31*(2), 33-37.

Camera, L. (2015, September 21). As test results trickle in, states still ditching Common Core. *U.S. News and World Report*. Retrieved from https://www.usnews.com

Carnoy. M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305-331.

Cavanagh, S. (2015, July 31). As McGraw-Hill Education leaves state testing, market thrives for classroom assessments. *Education Week*. Retrieved from http://www.edweek.org

Chandler, M. A. (2014, April 5). Virginia students will take fewer Standards of Learning tests next year. *The Reporter*. Retrieved from http://www.thereporteronline.com

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. Retrieved from http://cs.colorado.edu/~jbg/docs/nips2009-rtl.pdf

Chen, Y., Yu, B., Zhang, X., Yu, Y. (2016, April). *Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals*. Paper presented at the International Conference on Learning Analytics and Knowledge, Edinburgh, U.K.

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science, 1*(10), 103-126.

Chubb, J. E. (1988). Why the current wave of school reform will fail. *The Public Interest, 90*, 28-49.

Cizek, G. J. (2000). Pockets of resistance in the assessment revolution. *Educational Measurement: Issues and Practice, 19*(2), 16-23.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19-27.

Cizek, G. J. (2005). High-stakes testing: Context, characteristics, critiques, and consequences. In R. Phelps (Ed.), *Defending standardized testing* (pp. 23-54). Mahwah, NJ: Lawrence Erlbaum.

Clark, W. R., & Golder, M. (2015). Big data, causal inference, and formal theory: Contradictory trends in political science? *PS: Political Science & Politics, 48*(1), 65-70.

Clinton, W. J. (1997). Address before a joint session of the Congress on the state of the union. The American Presidency Project. Retrieved from http://www.presidency.ucsb.edu

Clinton, W. J. (1999). Remarks on proposed legislation to reauthorize the Elementary and Secondary Education Act. The American Presidency Project. Retrieved from http://www.presidency.ucsb.edu

Cohen-Vogel, L. (2011). "Staffing to the test": Are today's school personnel practices evidence based? *Educational Evaluation and Policy Analysis, 33*(4), 483-505.

Cohen-Vogel, L., & Rutledge, S. (2009). The pushes and pulls of new localism: School-level instructional arrangements, instructional resources, and family-community partnerships. *Yearbook of the National Society for the Study of Education,*108(1), 70-103

College Board. (n.d.). *ACCUPLACER*. Retrieved from https://accuplacer.collegeboard.org/professionals

Croft, S. J., Roberts, M. A., & Stenhouse, V. L. (2016). The perfect storm of education reform: High-stakes testing and teacher evaluation. *Social Justice, 42*(1). 70-92.

Dearing, J. W., & Rogers, E. M. (1996). *Agenda-setting.* Thousand Oaks, CA: Sage.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Denham, B. E. (2014). Intermedia attribute agenda setting in the New York Times: The case of animal abuse in U.S. horse racing. *Journalism and Mass Communication Quarterly, 91*(1), 17-37.

Development Process. (n.d.). Common Core State Standards initiative. Retrieved from http://www.corestandards.org/about-the-standards/development-process/

Dillon, S. (2004, January 2) Some school districts challenge Bush's signature education law. *The New York Times*. Retrieved from http://www.nytimes.com

Dillon, S. (2006a, March 26). Schools cut back subjects to push reading and math. *New York Times*. Retrieved from http://www.nytimes.com

Dillon, S. (2006b, November 20). Schools slow in closing gaps between races. *The New York Times*. Retrieved from http://www.nytimes.com

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*. Retrieved from http://bds.sagepub.com

DiMaggio, P., Nag, M., & Blei, D. M. (2013). Exploiting affinities between topic models and the sociological perspective on culture: Applications to newspaper coverage of U.S. government arts funding. *Poetics*, 41, 570–606.

Duncan, A. (2009, July 24). Education reform's moon shot. *The Washington Post*. Retrieved from http://www.washingtonpost.com

Dwyer, J. G. (2004). Introduction to symposium: School accountability and 'high stakes' testing. *Theory and Research in Education, 2*(3), 211-217.

Edelman, M. (1988). *Constructing the political spectacle*. Chicago, IL: University of Chicago Press.

Editorial Projects in Education. (n.d.) Mission and history. Retrieved from http://www.edweek.org

Education Commission of the States. (n.d.). State legislation: High school exit exams. Retrieved from http://www.ecs.org

Ehrenfreund, M. (2015, April 14). Why civil rights groups support standardized tests. *The Washington Post.* Retrieved from https://www.washingtonpost.com

EMC. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. Retrieved from http://www.emc.com

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51-58.

Entman, R. M. (2004). *Projections of power: Framing news, public opinion, and U.S. foreign policy*. Chicago, IL: University of Chicago Press.

Evers, W. F., & Walberg, H. J. (2004). *Testing student learning, evaluating teaching effectiveness*. Stanford, CA: Hoover Press.

Every Student Succeeds Act, 20 USC § 6301 (2015).

FairTest. (2016, May 9). More than 670,000 Refused Tests in 2015. Retrieved from http://www.fairtest.org

Felton, E. (2015a, August 18). Poll raises questions about the breadth of the opt-out movement. *Hechinger Report*. http://hechingerreport.org

Felton, E. (2015b, September 25). Experts predict the opt-out movement will get some of what it wants. *Hechinger Report*. http://hechingerreport.org

Fiedler, R. (2014, April 1). Parents claim victory as Waco ISD allows opt-out option for STAAR test. The *Baylor Lariat*. Retrieved from http://baylorlariat.com

Fine, M., & Jaffe-Walter, R. (2007). Swimming: On oxygen, resistance, and possibility for immigrant youth under siege. *Anthropology and Education Quarterly, 38*(1), 76-96.

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998) Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20*(2), 95-113.

Foorman, B. R., Fletcher, J. M., & Francis, D. J. (2004) Early reading assessment. In W. Evers & H. Walberg (Eds.), *Testing student learning, evaluating teaching effectiveness* (pp. 81-125). Stanford: Hoover Institution Press.

Friedman, D. (2012, March 16). The rise of big data. *New York Academy of Sciences Magazine*. Retrieved from http://www.nyas.org

Fryer, R. G. (2010). *Financial incentives and student achievement: Evidence from randomized trials* (NBER Working Paper No. 15898). Cambridge, MA: National Bureau of Economic Research.

Fuller, M. B. (2014). A history of financial aid to students. *Journal of Student Financial Aid, 44*(1), 42-68.

Gabriel, T. (2010, June 10). Under pressure, teachers tamper with tests. *New York Times*. Retrieved from http://www.nytimes.com

Gamson, W. A. (1992). *Talking politics*. Cambridge, U.K.: Cambridge University Press.

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology, 95*(1), 1-37.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. IDC Iview. Retrieved from http://www.emc.com

Gardner, W. (2014, August 13). Atlanta cheating scandal lessons. *Education Week*. Retrieved from http://edweek.org/

Gewertz, C. (2010, May 10). Study of reading programs finds little proof of gains in student comprehension. *Education Week*. Retrieved from http://edweek.org/

Gewertz, C. (2014, August 21). U.S. Ed. Sec. Duncan: Too much testing costs teachers and students 'precious time'. *Education Week*. Retrieved from http://edweek.org/

Ghassemi, M., Naumann, T., Joshi, R., & Rumshisky, A. (2012). Topic models for mortality modeling in intensive care units. Retrieved from http://people.cs.pitt.edu/~milos/icml_clinicaldata_2012/Papers/Poster_GhassemiNaumann_ICML_Clinical_2012.pdf

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York: Peter Lang Publishing.

Golan, G. (2006). Inter-media agenda setting and global news coverage: Assessing the influence of the New York Times on three network television evening news programs. *Journalism Studies, 7*(2), 323-333.

Gould, S. J. (1996). *The mismeasure of man* (2nd ed.). New York, NY: WW Norton.

Graduation test update. (2017, April). Retrieved from http://www.fairtest.org

Great Schools Partnership. (n.d.) *The glossary of education reform*. Retrieved from http://edglossary.org

Greenhouse, S. (2001, May 4). Union signals softer stance on merit pay. *New York Times*. Retrieved from http://www.nytimes.com

Gregory, R. J. (2013). *Psychological testing: History, principles, and applications* (2nd ed.). London: Pearson.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis, 18*, 1-35.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promises and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 1-31.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. Haertel (Eds.), *Uses and misuses of data in accountability testing* (pp. 1-34). Malden, MA: Blackwell.

Harris, E. A. (2015, March 1). As Common Core testing is ushered in, parents and students opt out. *New York Times*. Retrieved from http://www.nytimes.com

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's great city schools: An inventory and preliminary analysis*. Washington, DC: Council of Great City Schools.

Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis, 30*(2), 75-110.

Herman, J. L., & Baker, E. L. (2009) Assessment policy: Making sense of the Babel. In G. Sykes, B. Schneider, D. Plank, & T. G. Ford (Eds.), *Handbook of education policy research* (pp. 176-190). New York, NY: Routledge.

Herman, J. L., & Golan, S. (1990). *Effects of standardized testing on teachers and learning: Another look*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 341 738).

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12*(4), 20-25.

Herman, J. L., & Haertel, E. H. (Eds.) (2005). *Uses and misuses of data in accountability testing*. Malden, MA: Blackwell.

Herman, J. L. (2008). Accountability and assessment: Is public interest in K-12 education being served? In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 211-231). New York, NY: Routledge.

Hernandez, J. C. (2014, April 9). New York City reducing role of tests in advancing students to next grade. *New York Times*. Retrieved from http://www.nytimes.com

Herszenhorn, D. M. (2005, September 23). Math scores statewide show gains in 4th grade. *New York Times*. Retrieved from http://www.nytimes.com

Heubert, J.P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Hoff, D. J. (2008, February 12). U.S. 'dashboards' offer data on state achievement. *Education Week*. Retrieved from http://edweek.org/

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Hout, M., & Elliott, S. W. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington, DC: National Research Council.

Hursh, D. (2008). The growth of high-stakes testing in the USA: Accountability, markets, and the decline in educational equality. In W. Harlen (Ed.), *Student assessment and testing, volume 4* (pp. 275-293). Los Angeles, CA: Sage.

Hyslop, A. (2014, July 15). The case against exit exams. New America Foundation. Retrieved from https://www.newamerica.org

Iyengar, S. (1991). *Is anyone responsible? How television frames political issues*. Chicago, IL: University of Chicago Press.

Iyengar, S. (1997). The effects of news on the audience: Overview. In S. Iyengar & R. Reeves (Eds.), Do the media govern? Politicians, voters, and reporters in America (pp. 211-216). Thousand Oaks, CA: Sage.

Iyengar, S., & Kinder, D. R. (1987). *News that matters: Television and American opinion.* Chicago, IL: University of Chicago Press.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*, 761-796.

Jacob, B. A. & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1), 226-244.

Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating* (NBER Working Paper No. 9413). Cambridge, MA: National Bureau of Economic Research.

Jacobsen, R. (2009). The voice of the people in education policy. In G. Sykes, B. Schneider, D. Plank, & T. G. Ford (Eds.), *Handbook of education policy research* (pp. 307-318). New York, NY: Routledge.

Jennings, J. (1998). *Why national standards and tests? Politics and the quest for better schools*. Thousand Oaks, CA: Sage.

Jennings, J. (2016). Fifty years of federal aid to school: Back into the future? *Education Law and Policy Review, 3*, 1-30.

Jennings, J., & Rentner, D. S. (2006). Ten big effects of the No Child Left Behind Act on public schools. *Phi Delta Kappan, 88*(2), 110-113.

Jo, Y., Loghmanpour, N., & Rose, C. (2015). *Time series analysis of nursing notes for mortality prediction via a state transition topic model*. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 1171-1180.

Jochim, A. E., & Jones, B. D. (2013). Issue politics in a polarized congress. *Political Research Quarterly, 66*(2), 352-369

Jockers, M. (2014) *Text analysis with R for students of literature*. Switzerland: Springer International.

Jockers, M. L., & Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing, 25*(2), 215-224.

Johnson, J. (2013). *Will it be on the test? A closer look at how leaders and parents think about accountability in public schools*. New York, NY: Kettering Foundation and Public Agenda.

Jones, B. D., & Wolfe, M. (2010). Public policy and the mass media: An information processing approach. In S. Koch-Baumgarten & K. Voltmer (Eds.), *Public policy and mass media* (pp. 17-43). New York, NY: Routledge.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Katsiyannis, A., Zhang, D., Ryan, J. B., & Jones, J. (2007). High-stakes testing and students with disabilities. *Journal of Disability Policy Studies, 18*(3), 160-167.

Kean, M. H. (2003). Educational assessment in a reform context. In J. Wall & G. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 335-347). Greensboro, NC: Caps Press.

Keeling, D. (2014, February, 13). Fixing the culture of testing. TNTP Blog. http://tntp.org/

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science, 331*(6018), 719-721.

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science." *PS: Political Science and Politics, 47*(1), 165-172.

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review, 107*(2), 1-18.

Klein, A. (2016). The Every Student Succeeds Act: An ESSA overview. *Education Week*. http://www.edweek.org.

Klein, J. (2011, June). The failure of American schools. *The Atlantic*. Retrieved from http://www.theatlantic.com

Kober, N. (2015). *Knowing the score: The who, what, and why of testing*. Washington, DC: Center on Education Policy.

Krieg, J. M. (2008). Are students left behind? The distributional effects of the No Child Left Behind Act. *Education Finance and Policy, 3*(2), 250-281.

Ladd, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426-450.

Lauen, D., & Gaddis, M. (2012). Shining a light or fumbling in the dark? The effects of NCLB's subgroup-specific accountability pressure on student performance." *Educational Evaluation and Policy Analysis 34*(2), 185-208.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311-331.

Leadership Conference on Human and Civil Rights. (2015). Civil rights groups: "We oppose anti-testing efforts" [Press release]. Retrieved from http://www.civilrights.org/press/2015/anti-testing-efforts.html

Lee, J. K. (2007). The effect of the internet on homogeneity of the media agenda: A test of the fragmentation thesis. *Journalism and Mass Communication Quarterly, 84*(4), 745-760.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus, and Giroux.

Lindle. J. C. (2009). Assessment policy and politics of information. In G. Sykes, B. Schneider, D. Plank, & T. G. Ford (Eds.), *Handbook of education policy research* (pp. 319-332). New York, NY: Routledge.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L. (2008). Educational accountability systems. In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3-24). New York, NY: Routledge.

Lu, A. (2014, January 27). States reconsider Common Core test. *Huffington Post*. Retrieved from http://www.huffingtonpost.com

Lynch, M. (2015, August, 27). 10 reasons the U.S. education system is failing. *Education Week.* http://www.edweek.org

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education, 65*(3), 29-46.

Manna, P. (2006). *School's in: Federalism and the national education agenda*. Washington, DC: Georgetown University Press.

Matthes, J. (2009). What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism and Mass Communication Quarterly, 86*(2), 349-367.

Matthes, J. & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication, 58*(2), 258-279.

McCombs, M. (2004). *Setting the agenda: The mass media and public opinion*. Cambridge, MA: Polity Press.

McCombs, M., & Ghanem, S. I. (2001). The convergence of agenda setting and framing. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives of media and our understanding of the social world* (pp. 67-82). Mahwah, NJ: Erlbaum.

McDonnell, L. M. (1994). Assessment policy as persuasion and regulation. *American Journal of Education, 102*(4), 394-420.

McDonnell, L. M. (1997). *The politics of state testing: Implementing new student assessments* (CSE Technical Report 424). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

McDonnell, L. M. (2005). Assessment and accountability from the policymaker's perspective. *Yearbook of the National Society for the Study of Education, 104*(2), 35-54.

McDonnell, L. M. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education, 80*(2), 19-38.

McDonnell, L. M. (2008). The politics of educational accountability: Can the clock be turned back? In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 47-67). New York, NY: Routledge.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics, 41*, 607-625.

McIntosh, S. (2012). *State high school exit exams: A policy in transition*. Washington, DC: George Washington University, Center on Education Policy.

McLaughlin. M. (1987). Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis, 9*(2), 171-178.

Medina, J. (2010, October 10). On New York school tests, warning signs ignored. *New York Times*. Retrieved from http://www.nytimes.com

Mehrens, W. A. (2004). Using performance assessment for accountability purposes. In W. Evers & H. Walberg (Eds.), *Testing Student Learning, Evaluating Teaching Effectiveness* (pp. 221-242). Stanford, CA: Hoover Institution Press.

Mehta, J. (2015, Summer). Escaping the shadow: A Nation at Risk and its far-reaching influence. *American Educator*, 20-26.

Michaels, P. (2015, July 22). Civil rights groups split over opting out of standardized tests. *Texas Observer*. Retrieved from https://www.texasobserver.org

Miller, M. D. (2008). Data for school improvement and educational accountability: Reliability and validity in practice. In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 249-261). New York, NY: Routledge.

Miller, M. M. (1997). Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review, 15*(4), 367-378.

Miller, M. M., & Riechert, B. P. (2001). The spiral of opportunity and frame resonance: Mapping the issue cycle in news and public discourse. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives of media and our understanding of the social world* (pp. 107–121). Mahwah, NJ: Erlbaum.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham, MA: Elsevier.

Mitchell, C. (2015, April 8). Convicted Atlanta educators draw empathy, condemnation. *Education Week*. Retrieved from http://www.edweek.org

Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015) "No! Formal theory, causal inference, and big data are not contradictory trends in political science." *PS: Political Science and Politics, 48*(1), 71-74.

Moses, M. S., & Nanna, M. J. (2007). The testing culture and the persistence of high stakes testing reforms. *Education and Culture, 23*(1), 55-72.

National Commission for Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Conference of State Legislatures. (2005). *Task force on No Child Left Behind: Final report*. Denver, CO: National Conference of State Legislatures.

National Education Association (2009). NEA's response to Race to the Top. [Press release]. Retrieved from http://nea.org

National Education Association. (2013). NEA President supports Seattle educators who refuse to give flawed standardized test [Press release]. Retrieved from http://nea.org

National Governors Association. (1986). *Time for results: The governors' 1991 report on education*. Washington, DC: National Governors Association.

Nelson, R. K. (2011, May 29). Of monsters, men — and topic modeling. *New York Times*. Retrieved from http://www.nytimes.com

Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology, 57*(6), 753–767.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved from http://epaa.asu.edu/

Nowlin, M. (2016). Modeling issue definitions using quantitative text analysis. *Policy Studies Journal, 44*(3), 309-331.

OBrien, A. (2016, January 28). 5 ways ESSA impacts standardized testing. *Edutopia*. Retrieved from https://www.edutopia.org

O'Connor, B., Bamman, D., Smith, N. A. (2011). *Computational text analysis for social science: Model assumptions and complexity*. Paper presented at Second NIPS Workshop on Computational Social Science and the Wisdom of Crowds.

Office of the Governor, State of Georgia. (2011). *Special investigation into test tampering in Atlanta's school system*. Retrieved from https://archive.org/stream/215252-special-investigation-into-test-tampering-in/

Olson, L. (2005, November 29). Shifts in state systems for gauging AYP seen as impeding analysis. *Education Week*. Retrieved from http://edweek.org/

Opfer, V. D. (2007). Developing a research agenda on the media and education. *Peabody Journal of Education, 82*(1), 166-177.

Orfield, G. (2016). A great federal retreat: The 2015 Every Student Succeeds Act. *Education Law and Policy Review, 3*, 273-288.

Page, B. I., & Shapiro, R. Y. (1983) Effects of public opinion on policy. *The American Political Science Review, 77*(1), 175-190.

Paige, R. (2003). Education in America: The complacency must end. *Education Resources Information Center*. Retrieved from http://files.eric.ed.gov

Perry, J., Judd, A. & Pell, M. (2012, March 25). Cheating our children: Suspicious school test scores across the nation. *Atlanta Journal-Constitution*. Retrieved from http://www.myajc.com

Phelps, R. P. (1998). The demand for standardized student testing. *Educational Measurement: Issues and Practice, 17*(3), 5-23.

Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ: Transaction Publishers.

Phelps, R. P. (Ed.). (2005). *Defending standardized testing*. Mahwah, NJ: Lawrence Erlbaum.

Pizmony-Levy, O. & Green Saraisky, N. (2016). *Who opts out and why? Results from a national survey on opting out of standardized tests.* Research Report. New York: Teachers College, Columbia University.

Porter, A. (n.d.). Rethinking the achievement gap. Retrieved from www.gse.upenn.edu

Price, V., Tewksbury, D., & Powers, E. (1997). Switching trains of thought: The impact of news frames on readers' cognitive responses. *Communication Research, 24*(5), 481-506.

Prior, M. (2003). Any good news in soft news? The impact of soft news preference on political knowledge. *Political Communication, 20*(2), 149-171.

Public Agenda. (2006). *Is support for standards and testing fading? Reality check 2006* (Issue no. 3). Retrieved from http://www.publicagenda.org

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Rado, D. (2014, May 11). Pushback in suburbs over state testing. *Chicago Tribune*. Retrieved from http://articles.chicagotribune.com

Rasmussen, S. (2015, March 11). Why the Smarter Balanced Common Core math test is fatally flawed. *EdSurge*. Retrieved from https://www.edsurge.com/

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy, 6(3),* 207–241

Reese, S. D. (2001). Framing public life: A bridging model for media research. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives of media and our understanding of the social world* (pp. 7-31). Mahwah, NJ: Erlbaum.

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Cambridge, MA: Harvard University Press.

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. M. (2015). Computer-assisted reading and discovery for student-generated text in massive open online courses. *Journal of Learning Analytics*, *2*(1), 156–184.

Rich, M. (2015, October 28). Nationwide test shows dip in students' math abilities. *New York Time*s. Retrieved from http://www.nytimes.com

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2015). *A model of text for experimentation in the social sciences*. Retrieved from http://scholar.princeton.edu

Roberts, M. E., Stewart, B. M., and Tingley, D. (2014). *stm: R package for structural topic models* [Technical report]. Cambridge, MA: Harvard University.

Rochefort, D. A., & Cobb, R. W. (1993). Problem definition, agenda access, and policy choice. *Policy Studies Journal, 21*(1), 56-71.

Rogers, E. W., Dearing, J. W., & Chang, S. (1991). AIDS in the 1980s: The agenda-setting process for a public issue. Journalism Monographs no. 126. Association for Education in Journalism and Mass Communication.

Romney, D., Stewart, B., & Tingley, D. (2015). Plain text? Transparency in computer-assisted text analysis. *Qualitative and Multi-Method Research, 13*(1), 32-38.

Rose, M., & Baumgartner, F. R. (2013). Framing the poor: Media coverage and U.S. poverty policy, 1960-2008. *The Policy Studies Journal, 41*(1), 22-53.

Rothman, R. (2011). *Something in common: The Common Core Standards and the next chapter in American education*. Cambridge, MA: Harvard Education Press.

Ryan, K. E. (2008). Fairness issues and educational accountability. In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 191-208). New York, NY: Routledge.

Ryan, K. E. & Shepard, L. A. (Eds.). (2008). *The future of test-based educational accountability*. New York, NY: Routledge.

Samuels, C. (2011, August 4). Cheating scandals intensify focus on test pressures. *Education Week*. Retrieved from http://edweek.org/

Sandham, J. L. (2000, February 2). Calif. schools get rankings based on tests. *Education Week.* Retrieved from http://edweek.org/

Sawchuk, S. (2009, August 25). NEA at odds with Obama team over 'Race to the Top' criteria. *Education Week*. Retrieved from http://edweek.org/

Scheufele, D. A. (1999) Framing as a theory of media effects. *Journal of Communication,* 49(1), 103-122.

Schneider, A., & Ingram, H. (1990). Behavioral assumptions of policy tools. *The Journal of Politics, 52*(2), 510-529.

Schweig, J. (2016, May 10). The opt-out reckoning. *U.S. News & World Report*. http://www.usnews.com/

Scott-Clayton, J. (2012). Do high-stakes placement exams predict college success? (CCRC Working Paper No. 41). New York, NY: Columbia University, Teachers College, Community College Research Center.

Shanahan, E. A., McBeth, M. K., Hathaway, P. L., & Arnell, R. J. (2008). Conduit or contributor? The role of media in policy change theory. *Policy Sciences, 41*(2), 115-138.

Shaw, D., & McCombs, M. (1977). *The emergence of American political issues.* St. Paul, MN: West.

Shepard, L. A., & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction*. Paper presented at the Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 337 468).

Siegel, H. (2004). High-stakes testing, educational aims and ideals, and responsible assessment. *Theory and Research in Education, 2*(3), 219-233.

Sjøberg, S. (2007). *PISA and 'real life challenges': Mission impossible?* Retrieved from http://folk.uio.no/sveinsj/Sjoberg-PISA-book-2007

Snow, D. A., & Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization. *Research in Social Movements, Conflicts and Change, 1*, 197-217.

Springer, M., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D. F.,…Stecher, B. M. (2012). Final report: Experimental evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Stack, M. (2006). Testing, testing, read all about it: Canadian press coverage of the PISA results. *Canadian Journal of Education, 29*(1), 49-69.

Stecher, B. (2004). Portfolio assessment and education reform. In W. Evers & H. Walberg (Eds.), *Testing Student Learning, Evaluating Teaching Effectiveness* (pp. 197-220). Stanford, CA: Hoover Institution Press.

Steele, C. M. (1997). A threat in the air: How stereotypes shape the intellectual identities and performance of women and African-Americans. *American Psychologist, 52*, 613-629.

Stobart, G., & Eggen, T. (2012). High-stakes testing - value, fairness, and consequences. *Assessment in Education: Principles, Policy, & Practice, 19*(1), 1-6.

Strauss, V. (2014, May 15). Pushback on standardized testing around the country getting stronger. *The Washington Post*. Retrieved from https://www.washingtonpost.com

Strauss, V. (2016, January, 31). The testing opt-out movement is growing, despite government efforts to kill it. *The Washington Post*. Retrieved from https://www.washingtonpost.com

Stullich, S., Eisner, E., & McCrary, J. (2008). *National assessment of Title I, final report: Volume I: Implementation*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Summers, J. (2014, May 7). Nation's report card shows stagnant scores for reading, math. *National Public Radio*. Retrieved from http://www.npr.org

Tanata, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43-50.

Tankard, J. W. (2001). An empirical approach to the study of media framing. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives of media and our understanding of the social world* (pp. 95–106). Mahwah, NJ: Erlbaum.

Teoh, M., Coggins, C., Guan, C., & Hiler, T. (2014). *The student and the stopwatch: How much time do American students spend on testing?* Boston, MA: Teach Plus.

Turner, D. (2011, July 16). 'Run like the mob': U.S. school cheating scandal details emerge. *NBC News*. Retrieved from http://www.nbcnews.com

Ujifusa, A. (2014a, May 6). State political rifts sap support for Common Core tests. *Education Week*. Retrieved from http://www.edweek.org

Ujifusa, A. (2014b, August 5). Standards persist amid controversy. *Education Week*. Retrieved from http://www.edweek.org

Ujifusa, A. (2016, January 14). Opt-out activists aim to build on momentum in states. *Education Week*. Retrieved from http://www.edweek.org

Underwood, T. (2012, April 7). Topic modeling made just simple enough [Blog post]. Retrieved from https://tedunderwood.com

U.S. Department of Education. (2013, August 29). *States granted waivers from No Child Left Behind allowed to reapply for renewal for 2014 and 2015 school years* [Press release]. Retrieved from https://www.ed.gov

U.S. Department of Education. (2015a). Fact sheet: Testing action plan [Press release]. Retrieved from http://www.ed.gov

U.S. Department of Education. (2015b). *The condition of education 2015*. Retrieved from www.nces.ed.gov

Viadero, D. (2003a, January 8). Reports find fault with high-stakes testing. *Education Week*. Retrieved from http://www.edweek.org

Viadero, D. (2003b, February 5). Researchers debate impact of tests. *Education Week*. Retrieved from http://www.edweek.org

Viadero, D. (2003c, April 16). Study finds higher gains in states with high-stakes tests. *Education Week*. Retrieved from http://www.edweek.org

Viadero, D. (2004, January, 28). Study offers mixed results on impact of high-stakes tests. *Education Week*. Retrieved from http://www.edweek.org

Vockell, E. L. (1993). Why schools fail and what we can do about it. *The Clearing House*, *66*(4), 200-205.

Voltmer, K., & Koch-Baumgarten, S. (2010). Introduction: Mass media and public policy - is there a link? In S. Koch-Baumgarten & K. Voltmer (Eds.), *Public policy and mass media* (pp. 1-13). New York, NY: Routledge.

Wachen, J. (2014, October). *Testing as a mechanism for excellence and equity: A solution becomes a problem.* Paper presented at the meeting of the Northeastern Educational Research Association, Trumbull, CT.

Walgrave, S., & Van Aelst, P. (2006). The contingency of the mass media's political agenda setting power: Toward a preliminary theory. *Journal of Communication, 56*, 88-109.

Warmington, P., & Murphy, R. (2004). Could do better? Media depictions of UK educational assessment results, *Journal of Education Policy, 19*(3), 285-299.

Warren, J. R., Jenkins, K. N., & Kulick, R. B. (2006). High school exit examinations and the state-level completion and GED rates, 1975-2002. *Educational Evaluation and Policy Analysis*, *28*(2), 131-152.

West, D. M., Whitehurst, G. J., & Dionne, E. J. (2009). *Invisible: 1.4 percent coverage for education is not enough*. Washington, DC: Brookings Institution.

Wiggins, E. L. (2001) Frames of conviction: The intersection of social frameworks and standards of appraisal in letters to the editor regarding a lesbian commitment ceremony. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives of media and our understanding of the social world* (pp. 207-214). Mahwah, NJ: Erlbaum.

Wolfe, M., Jones, B. D., & Baumgartner, F. R. (2013). A failure to communicate: Agenda setting in media and policy studies. *Political Communication, 30*(2), 175-192.

Wong, M., Cook, T. D., & Steiner, P. (2009). *No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series* (WP-09-11). Evanston, IL: Northwestern Institute for Policy Research.

Workman, E. (2014). *Third-grade reading policies*. Denver, CO: Education Commission of the States.

Zernike, K. (2001, April 13) In high scoring Scarsdale, a revolt against state tests. *New York Times.* Retrieved from http://www.nytimes.com